THE UNIVERSITY OF QUEENSLAND

AUSTRALIA

Population diversity and genomics of the fungal pathogen of canola *Leptosphaeria maculans*

Dhwani Apurva Patel

Bachelor of Biotechnology (Hons)

A thesis submitted for the degree of Doctor of Philosophy at

The University of Queensland in 2015

School of Agriculture and Food Sciences

Abstract

Canola (*Brassica napus*) is an important agricultural crop in Australia. Its value as an agricultural commodity has dramatically increased over the past few years to >$2000 million (AUD) in 2015. However, the production of this prosperous crop is threatened by up to 34 different pests, including fungal, bacterial and viral pathogens. Diseases caused by these pathogens lead to substantial crop losses that collectively amount to $130 million (AUD) each year. *Leptosphaeria maculans* (blackleg) is a ubiquitous ascomycete fungus that is the major causal agent of disease on canola plants. It is called blackleg due to its necrotrophic effects that cause stem canker at the base of the stem during the last stages of infection. When canola was first introduced to Australia as an agricultural crop in the 1970s, blackleg disease led to almost 90% crop losses, threatening to drive the nascent canola industry in Australia to the ground. In recent years, crop losses still amount to $76.6 million (AUD) per annum. Therefore, it is imperative to devise methods to control the devastating effects of this pathogen to protect this economically significant crop. In order to do so, we must decipher the genomic content that drives this pathogen to cause large-scale infections in the field.

Blackleg disease is a major concern not only in Australia, but also in other parts of the world such as Europe and Canada. Over the past few years, scientists have made significant advances in decoding the genome of this pathogen. It has been established that *L. maculans* interacts in a gene-for-gene manner with its host. It is composed of disease-causing avirulence (*Avr*) genes that are specifically recognised by the corresponding resistance (*R*) gene in the host plant. The specific location and characterisation of the *Avr* genes *AvrLm1*, *AvrLm6, AvrLm4-7, AvrLm11, AvrLmJ1, AvrLm2* and *AvrLm3* has been determined. Furthermore, the reference genome for this pathogen was sequenced in 2011. The reference genome based on the strain v23.1.3 is 45.12 Mb and is scaffolded onto 76 supercontigs. It was reported to be bipartite in nature, composed of AT and GC-rich blocks and riddled with truncated copies of transposable elements (TEs) that were affected by Repeat-Induced Point (RIP) mutations. The availability of a reference genome has greatly assisted in designing bioinformatics tools to analyse the genomic content of this pathogen. Here we contribute to the growing pool of knowledge of blackleg genomics.

This thesis begins by shedding light on the population diversity of this pathogen in Australia. We used Single Nucleotide Polymorphisms (SNPs) and sampled a large variety of isolates from different canola-growing regions across Australia. We were able to conclude that the Australian blackleg population is panmictic in nature, where members of the population interact randomly with one another, leading to high diversity via sexual reproduction. Increased genetic diversity amongst

the population contributes significantly to increase the evolutionary potential of this pathogen that empowers it to overcome host resistance swiftly. Some clonal subpopulations were also observed, segregated by location of sampling. This indicated that isolates may be affected by different conditions such as temperature and rainfall that may promote asexual reproduction within subpopulations. The panmictic nature of the population was supported by the results of linkage disequilibrium and principle component analyses. The former displayed no association amongst markers and the latter, no groupings based on parameters associated with samples such as state sample was obtained from, stubble cultivar isolated from, stubbled resistance source or year of isolation.

Next we moved to study at the genomic level of the pathogen and conducted a gene loss analysis on nine isolates, which belonged to the differential set for which we possessed phenotypic information. This experiment was based on the hypothesis that *Avr* genes are deleted or lost in blackleg to avoid host *R* gene detection and confer virulence. We aimed to discover novel avirulence effectors through this method. Our results supported our previous findings in that conservation or loss of genomic content in the samples was not associated with any particular parameter. Two supercontigs, SC_13 and SC_12 displayed maximum and clustered gene loss in the dataset. Of these, SC_13 gene loss was annotated and found to be a part of a secondary metabolite cluster. Secondary metabolite clusters are disposable in nature, which explained the presence and absence of that set of genes amongst the nine isolates. Genes lost on SC_12 were concluded to be a part of a 'volatile' group of genes whose absence does not affect the fitness of the pathogen. There was also functional redundancy for genes that were lost, as results of infection studies on susceptible cultivars could not be correlated to loss of any particular gene. Other aspects such as effects of TEs and RIP mutations also play a role in the loss of genes. This study highlighted regions of the genome under selection pressure and the diversity in genomic content in isolates of blackleg.

Data obtained from other bioinformatics analysis such as SNP prediction and Presence/Absence Variation (PAV) was pooled with gene loss data to form a mega-resource to fill gaps in information about the genome. Firstly, this enabled us to identify three novel avirulence effector candidates A, B and C. Of these, gene C was a candidate for *AvrLm5*. All these displayed the common traits shared by effectors such as encoding small secreted proteins (SSPs), being upregulated during infection and located in an AT-rich region that contains no other predicted genes. Candidates A and B also displayed a high degree of conservation within a 12 amino acid motif, a feature that is uncommon amongst blackleg effectors. Both candidates also shared partial homology at the same motif with *AvrLmJ1*. One candidate was not annotated in the reference sequence but high LD in the

region of interest on SC_13, gene loss associated with *AvrLm8* virulence profile on cultivars and results from SignalP and Open Reading Frame analyses, led us to conclude the high probability of a novel effector being present in that region.

The experiments detailed in this thesis confirmed the presence of high genetic and genomic diversity in isolates of blackleg that enable it to overcome host selection pressure and rapidly cause infection. They also opened up several intriguing avenues of results that warrant further study. In this manner our results contributed to the pool of knowledge about blackleg. Information about the inner workings of this pathogen will greatly assist in the design of more robust, disease-resistant cultivars and preventing future crop losses in Australia.

Declaration by author

This thesis is composed of my original work, and contains no material previously published or written by another person except where due reference has been made in the text. I have clearly stated the contribution by others to jointly-authored works that I have included in my thesis.

I have clearly stated the contribution of others to my thesis as a whole, including statistical assistance, survey design, data analysis, significant technical procedures, professional editorial advice, and any other original research work used or reported in my thesis. The content of my thesis is the result of work I have carried out since the commencement of my research higher degree candidature and does not include a substantial part of work that has been submitted to qualify for the award of any other degree or diploma in any university or other tertiary institution. I have clearly stated which parts of my thesis, if any, have been submitted to qualify for another award.

I acknowledge that an electronic copy of my thesis must be lodged with the University Library and, subject to the policy and procedures of The University of Queensland, the thesis be made available for research and study in accordance with the Copyright Act 1968 unless a period of embargo has been approved by the Dean of the Graduate School.

I acknowledge that copyright of all material contained in my thesis resides with the copyright holder(s) of that material. Where appropriate I have obtained copyright permission from the copyright holder to reproduce material in this thesis.

Publications during candidature

Journal publications

Patel, D.A., Zander, M., Van de Wouw, A.P., Mason, A.S., Edwards, D. and Batley, J. (2015) Population diversity of *Leptosphaeria maculans* in Australia. *International Journal of Biology* 7, 18-36.

Zander, M., Alamery, S., Lorberg, S.K., **Patel, D.A**., Tollenaere, R., Edwards, D. and Batley, J. (2015) Understanding Effector-Triggered Immunity of Plants to Pathogens: A Genomics Approach. *Plant Biotechnology Journal* (Submitted)

Golicz, A.A., Martinez, P.A., Zander, M., **Patel, D.A**., Van De Wouw, A.P., Visendi, P., Fitzgerald, T.L., Edwards, D. and Batley, J. (2015) Gene loss in the fungal canola pathogen *Leptosphaeria maculans*. *Functional & Integrative Genomics* 15, 189-196.

Dalton-Morgan J., Hayward A., Alamery S., Tollenaere R., Mason A.S., Campbell E., **Patel D**., Lorenc M.T., Yi B., Long Y., Meng J., Raman R., Raman H., Lawley C., Edwards D. and Batley J. (2014) A high-throughput SNP array in the amphidiploid species *Brassica napus* shows diversity in resistance genes. *Functional & Integrative Genomics*. 14(4). 643-55.

Zander M.**, Patel D.A.**, Van de Wouw A.P., Lai K, Lorenc M.T., Campbell E., Hayward A., Edwards D., Raman H. and Batley J. (2013) Identifying genetic diversity of avirulence genes in *Leptosphaeria maculans* using whole genome sequencing. *Functional & Integrative Genomics*, 13(3), 295-308.

Book Chapter

Patel, D.A., Zander, M., Dalton-Morgan, J. and Batley, J. (2015) Advances in plant genotyping: where the future will take us. In: Plant Genotyping (Batley, J. ed.) pp. 1-11. Springer New York.

Conference abstracts/Oral presentations

Oral presentations

Patel D.A., Zander M., Golicz A.A., Chan K., Edwards D., Batley J. (Dec 2013) Understanding genome structure and diversity of *Leptosphaeria maculans* 2nd December 2013. Stromlo Molecular Plant Pathology meeting, Canberra.

Patel D.A., Zander M., Golicz A.A., Martinez P.A., Van de Wouw A., Hayashi S., Lorberg, Edwards D., Batley J. (Jan 2015) Patterns of Gene Loss in Population Genomic Data - Identifying Disposable Genes. 10th January 2015. Plant and Animal Genome conference XXIII, San Diego, California, USA.

Poster presentations

Patel D.A., Zander M., Van de Wouw A., Hayashi S., Lorberg S.K., Golicz A.A., Martinez P.A., S., Lee J., Hurgobin B., Chan K., Edwards D and Batley J. (Jan 2015) Using Gene loss and Presence/Absence Variation (PAV) analysis in the plant pathogen *Leptosphaeria maculans*. Plant and Animal Genome Conference XXIII, San Diego, California, USA.

Zander, M., **Patel, D.A**., Van de Wouw, A., Lai, K., Lorenc, M.T., Hayward, A.C., Golicz, A., Chan, K., Bayer, P., Edwards, D., and Batley, J. (Jan 2014) Understanding genome structure and diversity of the plant pathogen, Leptosphaeria maculans. Plant and Animal Genome Conference XXII, San Diego, California, USA.

Zander, M., **Patel, D.A**., Van de Wouw, A., Lai, K., Lorenc, M.T., Hayward, A.C., Golicz, A., Chan, K., Bayer, P., Edwards, D., and Batley, J. (March 2014) Understanding Genome Structure and Diversity of the Plant Pathogen, Leptosphaeria maculans. Crucifer Genetics Workshop and Brassica2014 Conference, Wuhan, China.

Publications included in this thesis

Patel, D.A., Zander, M., Van de Wouw, A.P., Mason, A.S., Edwards, D. and Batley, J. (2015) Population diversity of *Leptosphaeria maculans* in Australia. *International Journal of Biology* 7, 18-36.

(Incorporated into Chapter 3 of this thesis)

Contributor	Statement of contribution
Dhwani Apurva Patel	Designed experiments (60%) Conducted experiments (80%) Wrote the paper (90%)
Manuel Zander	Conducted experiments (20%) Wrote the paper (10%)
Angela P Van de Wouw	Edited paper (10%)
Annaliese S Mason	Designed experiments (10%) Edited paper (10%)
David Edwards	Designed experiments (10%) Edited paper (10%)
Jacqueline Batley	Designed experiments (20%) Edited paper (70%)

Golicz, A.A., Martinez, P.A., Zander, M., **Patel, D.A**., Van De Wouw, A.P., Visendi, P., Fitzgerald, T.L., Edwards, D. and Batley, J. (2015) Gene loss in the fungal canola pathogen *Leptosphaeria maculans*. *Functional & Integrative Genomics* 15, 189-196.

(Incorporated into Chapter 4 of this thesis)

Contributor	Statement of contribution
Agnieszka A Golicz	Designed experiments (60%) Conducted experiments (50%) Wrote the paper (90%)
Paula A Martinez	Conducted experiments (10%) Wrote the paper (10%)
Manuel Zander	Conducted experiments (10%)
Dhwani Apurva Patel	Conducted experiments (20%)
Angela P Van de Wouw	Edited paper (10%)
Paul Visendi	Conducted experiments (10%)
Timothy Fitzgerald	Edited paper (10%)
David Edwards	Designed experiments (20%) Edited paper (10%)
Jacqueline Batley	Designed experiments (20%) Edited paper (70%)

Contributions by others to the thesis

Chapter 3

Illumina GoldenGate analysis was conducted by Dr Manuel Zander as published in Zander et al (2013). Mr Philipp Bayer provided bioinformatics support for Manhattan plots. Dr Annaliese Mason provided help with R coding.

Chapter 4

Dr Manuel Zander, Dr Satomi Hayashi and Ms Sarah Lorberg aided in the isolation of fungal DNA and preparation of libraries for whole genome sequencing. Ms Agnieszka Golicz and Ms Paula Martinez provided bioinformatics support for SGSGeneLoss analysis. Ms Hua Yang assisted in conducting the PCRs.

Chapter 5

Dr Manuel Zander, Dr Satomi Hayashi and Ms Sarah Lorberg aided in the isolation of fungal DNA and preparation of libraries for whole genome sequencing. Dr Manuel Zander performed preliminary analysis for PAV and SNP data, as written in Zander et al (2015). The PAKAP pipeline was developed by Dr Kenneth Chan, Ms Jenny Lee and Ms Bhavna Hurgobin, who also provided bioinformatics support for the same. Mr Philipp Bayer provided bioinformatics support for RNA seq data download.

All chapters

Prof Jacqueline Batley provided ongoing assistance in experimental design and editing of the thesis. Prof David Edwards provided ongoing support for bioinformatics and editing of the thesis.

Statement of parts of the thesis submitted to qualify for the award of another degree

None

Acknowledgements

I graciously acknowledge the University of Queensland for providing me with access to world-class research facilities and opportunities, afforded to fortune students of this prestigious university. UQ has not only provided me with excellent learning but also fond memories from the past eight years that I will always cherish. I would also like to extend my thanks to the Australian Research Council, the School of Agriculture and Food Sciences and the Grains Research and Development Corporation for the scholarships and grants that made this PhD an enhanced learning experience.

Thank you to all the members of the Batley lab and CILR. The atmosphere of camaraderie, helpfulness and professionalism, constant in both labs, has not only taught me a lot but also made these last few years enjoyable. I am also grateful to Prof David Edwards and his bioinformatics group and Dr Annaliese Mason, for their support and advice.

Special thanks must be afforded to certain lab members, past and present. Alice and Emma who guided me through my first foray in scientific research and taught me excellent foundation skills. Satomi for her ever-available helpfulness and Sarah for her friendship and for always being there whenever I needed help in and outside the lab. Also, Alina for her boundless help and patience at being asked and answering the same questions and Reece, for his sarcasm, tremendous support, help and friendship throughout Honours and PhD.

Thank you Manny, firstly, for introducing me to this wonderful lab. I couldn't begin to describe how much your friendship and the support you have given me in various ways over the last six years, means to me. You're a gem of a person and I'm so grateful to have you as a close friend.

A big thank you to Canna and Shraddha, for making me feel like I was never far away from home, for being amazing friends, for all the support and for always being there for me. Also, Dula and Kavi, for countless chats, advice and friendship.

Thank you to everyone else who has been a part of this journey at some point or the other, my Australian family and thank you to all my friends scattered around the globe! Your love and support is omnipresent. Especially Devina, for being the one I can rely on always. I am truly lucky to be surrounded by such wonderful friends.

Next, my deepest and most grateful thanks go to Prof Jacqueline Batley, my supervisor. From supervising me during my third year research project, honours and this PhD, she has always

extended above and beyond to provide constant advice, support and encouragement. Your effort and confidence in my abilities has made this PhD a possibility and I am extremely fortunate to have you as my supervisor. Thank you for your understanding, patience, guidance and mentorship. You are a true role-model and I will always look up to you with respect.

Brij, you're my rock. Thank you for your patience, understanding, love and serene presence. You make life easy, even from so far away, I couldn't fathom doing this without you. No obstacle seems unsurpassable and no goal seems unattainable, with you by my side.

I am extremely grateful for being a part of an incredible family. Thank you Ba and Dada, I hope you're smiling down on me as I fulfil your wish to see me as a 'Dr'. Your values helped create a loving, supporting family without whom I am certain this journey would not have been easy. Thank you 'KK' for the joyous times, laughter, happiness and overflowing love, it radiates like rays of sunshine into my life. Especially Niti and Devanshi, who are there for me no matter what, in more ways than I can list; my support that I can rely on with my eyes closed.

Lastly but most importantly I'd like to thank my parents, Apurva and Mona. The vitality of your presence in my life is indescribable. Thank you for your unwavering support and encouragement. It has helped me reach new heights at every milestone in life. Your love and the values you instilled in me have made me who I am today and I would not be where I am without you. I am truly, truly blessed.

I dedicate this thesis to the both of you.

Keywords

Blackleg, gene loss, SNPs, goldengate, population diversity, next generation sequencing, brassica, genomics, presence absence variation

Australian and New Zealand Standard Research Classifications (ANZSRC)

ANZSRC code: 060408, Genomics, 30%
ANZSRC code: 060411, Population, Ecological and Evolutionary Genetics 30%
ANZSRC code: 060704, Plant Pathology, 40%

Fields of Research (FoR) Classification

FoR code: 0604 Genetics, 60%
FoR code: 0605 Microbiology, 40%

Table of Contents

List of Figures

List of Tables

List of abbreviations

Aa	Amino Acid
AT	Adenine Thymine
AFLP	Amplified Fragment Length Polymorphisms
AvrLm	Avirulent towards *Rlm* gene
Bp	Base pair
CDC	Conditionally Dispensable Chromosome
CDS	Coding Sequence of gene
Dpi	Days Post Inoculation
ETI	Effector-Triggered Immunity
GC	Guanine Cytosine
LD	Linkage Disequilibrium
LRR	Leucine Rich Repeat
Mb	Mega base (1 million bp)
MP	Mate Pair
NBS	Nucleotide Binding Site
NGS	Next Generation Sequencing
PAMP	Pathogen Associated Molecular Patterns
PAV	Presence/Absence Variation
PCA	Principle Component Analysis
PE	Paired End
PIC	Polymorphism Information Content
PTI	PAMP Triggered Immunity
RFLP	Restricted Fragment Length Polymorphisms
RIP	Repeat-Induced Point mutation
Rlm	Resistance to *Leptosphaeria maculans* gene
RLP	Receptor-Like Proteins
SC	SuperContig
SNP	Single Nucleotide Polymorphism
SSP	Small Secreted Protein
TE	Transposable element

1 Introduction

1.1 *Brassica napus*-Canola

Brassica napus (canola/rapeseed/oilseed rape) production in Australia has increased dramatically over the past few years, making this food crop very significant to the Australian economy. Canola accounts for more than ninety percent of the 3-4 tonnes of Australian oilseed production (Holden, 2012). According to the 2013-2014 issue of the "Value of Agricultural commodities produced, Australia" (ABS, 2015), national Canola oil production was valued at $2128.7 million (AUD). On an international scale, Australian export of canola accounts for 15-20% of the global trade, with Australia exporting over one million tonnes of seed to Japan, Pakistan, Europe and China (AOF, 2015). Within Australia, Western Australia (WA) has the highest production of canola followed by New South Wales (NSW), Victoria (Vic) and South Australia (SA) (ABS, 2015).

However, this prosperous crop is threatened by blackleg disease, also known as phoma disease or 'stem canker', caused by the fungus *Leptosphaeria maculans* (anamorph *Phoma lingam*) (West et al., 2001). Yield losses due to this fungus lead to monetary losses of an average $76.6 million (AUD) annually (Murray and Brennan, 2012). Blackleg disease is also prevalent in other major oilseed growing regions worldwide, including Europe and Canada.

1.2 *Leptosphaeria maculans*-Blackleg

The oldest record of *L. maculans* dates back to 1791 (Dilmaghani et al., 2009). Blackleg disease caused by this pathogen has been found on *B. napus, B. juncea*, *B. rapa* and *B. oleracea* (canola, vegetable and mustard crops) (West et al., 2001). As canola production has increased over the years, the severity of disease has also increased (Marcroft and Bluett, 2008; Kaur et al., 2009), and blackleg in Australia has been classified as 'highly virulent' (Purwantara et al., 1998; Purwantara et al., 2000). Australia suffered heavy crop losses, in some cases of almost 90%, in blackleg disease epidemics in the 1960s and 1970s, with a major epidemic event in 1972 (Bokor et al., 1975). The pathogen caused major devastation to the oilseed rape industry yet again in 2003 when polygenic resistant cultivar resistance failed, leading to 90% crop losses in the Eyre Peninsula, South Australia, within three years of release (Van de Wouw et al., 2010).

L. maculans survives in the stubble of infected plants, during which time it produces its primary inoculum via sexual reproduction, called ascospores (Figure 1.1). These are released from the pseudothecia on rainfall occurring during the late autumn sowing period in Australia (Howlett et al.,

2015). Ascospores can travel several kilometres via wind, infecting plants at both the adult and seedling stage through the leaf stomata. Next, the pathogen colonises intercellular spaces, eventually causing visible lesions in about 10 days (Howlett et al., 2015). The infection then spreads asymptomatically down the petiole and stem, leading to death of cortex cells that causes canker at the stem base. The pathogen undergoes asexual reproduction in leaf tissue, producing conidia, its asexual spores, at the site of infection (Rouxel and Balesdent, 2005) (Figure 1.1).

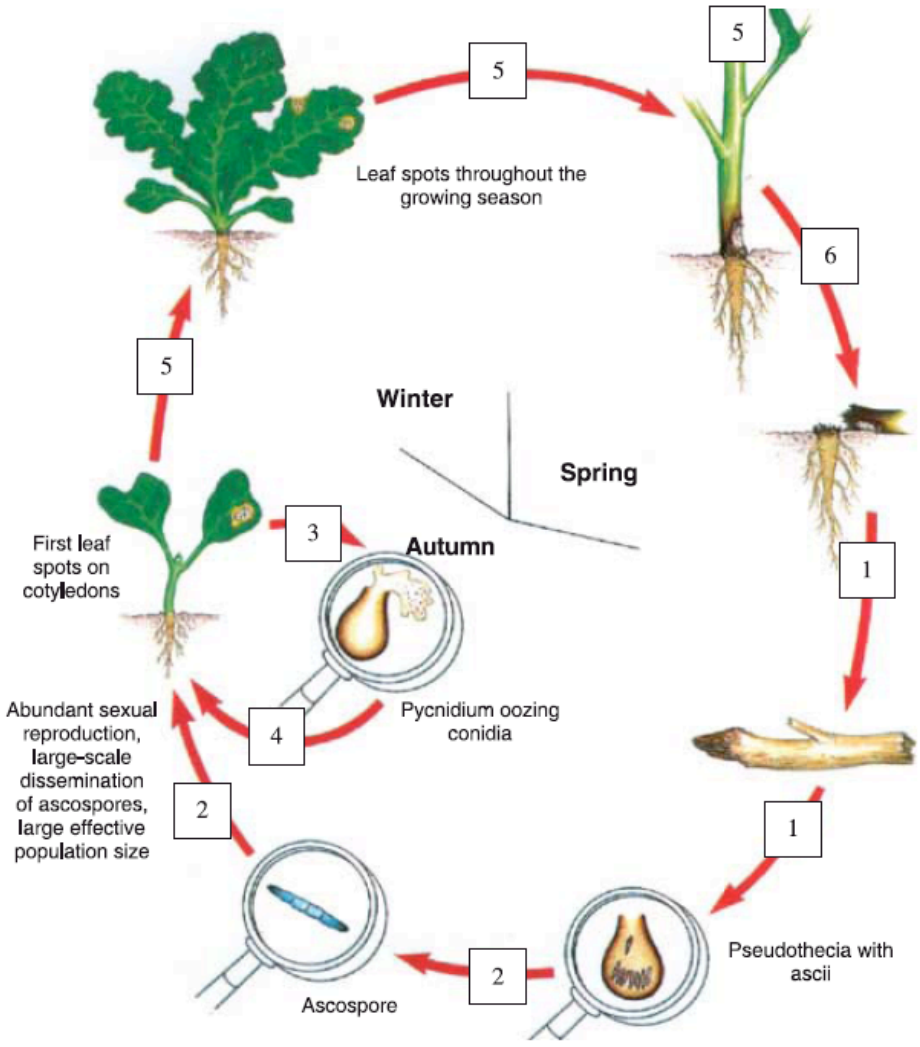

Figure 1.1 *L. maculans* life cycle

[(1) Survives as a saprobe on stubble (2) Sexual reproduction of ascospores (3) Infection through stomata; leaf spots as first sign of infection (4) Asexual reproduction of conidia (5) Spread of infection through plant vascular system (6) Necrosis of crown tissue or stem canker (Rouxel and Balesdent, 2005)]

1.2.1 Disease management

In order to prevent infection of new crops from blackleg, limiting their exposure to the previous year's crop stubble is imperative (Howlett et al., 2015). Current methods to conduct this include the use of fungicides (Khangura & Barbetti 2004), burning of infected stubble and rotating cultivars containing different sources of resistance (Salisbury et al. 1995). Marcroft et al. (2003) recommend that new canola cultivars should be sown at a distance of at least 500 m from old stubble. The best method to prevent crop losses, however, has been the sowing of disease-resistant cultivars, which is based on the interaction of effector and resistance proteins of the pathogen and plant (See section 1.3.1).

Despite these measures, there still remains a need for a more sustainable solution to prevent future crop losses. Understanding this pathogen at the genomic level and its population diversity at the genetic level will be vital in designing novel approaches to control blackleg infection in canola.

1.2.2 Population diversity

Basic techniques for estimating population diversity using molecular markers involve assessing how many alleles are present, determining how many individuals have unique (private) alleles and determining how much variation is present between individuals and within sub-populations. Similarity matrices, phylogenetic tree analyses, as well as principal component analysis, are a few tests which assess variance and relatedness in population samples. Other techniques used for estimating the genetic diversity in a population involve assessing how often alleles are recombined rather than inherited together. This is done by analysing the association among alleles from different loci (Xu, 2006). Alleles which are inherited together more often than predicted by chance are said to be in linkage disequilibrium (LD). Regions of LD can be due to evolutionary selective pressure for allele combinations or genome regions, or due to population stratification effects. In a panmictic population, less than 5% of loci are in significant LD (Xu, 2006).

In the past, blackleg population diversity has been studied using a variety of molecular marker systems. Diversity between Australian, North-American and European blackleg was classified using amplified fragment length polymorphisms (AFLPs) (Purwantara et al., 2000). Barrins et al. (2004) examined restricted fragment length polymorphisms (RFLP) and AFLPs in 194 Australian blackleg isolates, which revealed no genetic differences between the eastern and western *L. maculans* populations. However, Hayden et al. (2007) discovered east-west genetic differences using

minisatellite and microsatellite markers, attributing this finding to the presence of arid desert between the coasts. Travadon et al. (2011) used minisatellite markers to examine French blackleg isolates for indications of isolation by distance (IBD) and found that the French isolates operate as one panmictic population. A panmictic population is one in which members may reproduce with one another at random (Polk and Peek, 2010). On the other hand, Dilmaghani et al. (2012) proposed from minisatellite marker results that two genetically different populations of blackleg are present in Western Canada. Zander et al. (2013) conducted a small scale population diversity study using SNPs, whose successful results were used as a foundation for the large-scale study described in Patel et al. (2015) (Chapter 3).

1.3 Genomics of blackleg

The genome of *L. maculans* (v23.1.3 strain) was sequenced using a whole-genome shotgun strategy by Rouxel et al. (2011). The total genome of 45.12 Mb was scaffolded onto 76 SuperContigs (SCs). The data suggested the presence of 17-18 chromosomes, with ten of these corresponding to single SCs. Gene prediction analyses revealed 12,469 genes, of which nearly 85% were shown to be expressed (Rouxel et al., 2011). The *L. maculans* genome is made up of alternating homogenous blocks either rich in adenine and thymine (AT) or guanine and cytosine (GC), lending it its bipartite isochore-like nature. A total of 95% of predicted genes are located within GC-blocks whereas the other 5% are located within AT-blocks. Of that 5%, 122 genes encode probable small secreted proteins (SSPs). Other features of the blackleg genome include transposable elements (TEs), repeat-induced point mutations (RIPs) and a conditionally dispensable chromosome (CDC) which is the minichromosome SC_22 (Balesdent et al., 2013). The unique genomic makeup of the pathogen allows for rapid sequence diversification, and enables the fungus to evolve quickly against host resistance barriers.

1.3.1 Repeat-Induced Point (RIP) mutations

Selker et al. (1987) first described RIP mutations in *Neurospora crassa*. It was noticed that these mutations occurred before meiosis during dikaryon formation. Therefore, the acronym RIP was coined to stand for “rearrangement induced premeotically” but was later changed to mean “repeat-induced point” mutation (Hane et al., 2015). RIP is, simply put a mutating mechanism that affects repetitive sequences in the genome, the after effects of which may include epigenetic silencing through DNA methylation (Galagan and Selker, 2004). It is described as a defence mechanism that organisms employ to shield their genomes against debilitating effects of mobile genetic elements. RIP studies in *N. crassa* found that over 81% of repetitive sequences detected in the genome were

affected by RIP mutations (Galagan and Selker, 2004). RIP has been found to act preferentially at CpA dinucleotide sites (Cambareri et al., 1989), where the *RID1*-gene encoded C5-DNA-Methyltransferase (Jurkowski and Jeltsch, 2011) deaminates methylated cytosine to thymine (Galagan et al., 2003; Freitag et al., 2002).

Idnurm and Howlett (2003) first described the presence of the RIP mechanism in *L. maculans.* In blackleg, RIP mainly causes C to T and G to A mutations in transposable elements (TEs) and ribosomal DNA (Rouxel et al., 2011). It was noted that Pholy class 1 transposons degraded by RIP were more frequently located in pericentromeic regions (Attard et al., 2005). These mutations likely occur during sexual reproduction between isolates of these pathogen during its saprotrophic phase on the stubble (See Section 1.2). RIP is also known to affect genome evolution; however, this varies from species to species (Hane et al., 2015). In *N. crassa*, RIP prevents the prevalence of genes that share more than 80% similarity to each other (Selker, 1990; Kelkar et al., 2001; Galagan et al., 2003). This suggests that RIP limits genome evolution through gene family expansion via duplication. On the other hand, RIP proves beneficial in *L. maculans* because it leads to rapid evolution of diversified duplicated genes (Fudal et al., 2009; Van de Wouw et al., 2010). It is an ascertained method of divergence used by the pathogen to overcome host recognition of its disease-causing gene. RIP has also been documented to "leak" outside of repeat-regions into single-copy regions (Hane et al., 2015).

CDCs are also known to report a higher RIP intensity than the main chromosomes (Balesdent et al., 2013). They usually host genes that are not required for growth but play a role in other parts of the fungal life cycle, which usually confers a selective advantage to the pathogen (Balesdent et al., 2013). Mehrabi et al. (2011) note that *Alternaria alternata* carries host-selective toxins on its CDCs that enables it to cause infection on host plants. Similarly, in *Fusarium oxysporum*, Secreted In Xylem (SIX) genes located on the CDC, enabled it to adapt to tomato and confer virulence (Ma et al., 2010). Rouxel et al. (2011) also reported that the *L. maculans* CDC comprises a high number of TEs, which is consistent with the high degree of RIP mentioned earlier.

1.3.2 Transposable element (TE) composition

The *L. maculans* genome contains a high density of remnants of transposable elements at 32.5% (Rouxel et al., 2011; Grandaubert et al., 2014b). TEs can be divided into two major categories: Class I retrotransposons and Class II DNA transposons (Amselem et al., 2015). They are known to play a role in modifying the genomic structure of an organism via genome size expansion, gene

duplications, chromosomal rearrangements, inactivation of genes and gene loss (Amselem et al., 2015). In blackleg, 80% of detected TEs belong to the Class I category and they occur as clustered blocks, the majority of which are truncated mosaics (Rouxel et al., 2011). TEs are affected by RIP mutations and their mosaics are mainly found in AT-rich regions of the genome. The genomes of other pathogens such as *Phytophthora infestans* and *Blumeria graminis* have been known to undergo significant expansion due to the invasion of TEs and absence of mechanism that keep TE activity in check (Amselem et al., 2015). In *L. maculans* however, genome expansion via TE invasion has been controlled due to the activity of RIP that renders their transposition mechanism useless (Grandaubert et al., 2014b).

1.4 Effector triggered immunity

The term 'gene-for-gene resistance' was originally described by Flor in the 1940s (Flor, 1942). He proposed that plant resistance (*R*) genes form a one-on-one relationship with pathogen (*Avr*) genes. The term effector-triggered immunity (ETI) has evolved from the understanding of the gene-for-gene hypothesis. It is defined as the immune response that is generated in host plants upon recognition of a pathogen *Avr* gene by the plant *R* gene (Stuart et al., 2013) and is also known as a direct interaction. When the pathogen first infects the host, Pathogen Associated Molecular Patterns (PAMP)-triggered immunity (PTI) is initiated as plant pattern recognition receptors (PRRs) recognise PAMPs extracellularly (Stotz et al., 2014). In plants that are resistant to the pathogen, ETI is triggered leading to a hypersensitive response often causing localised cell death (Stotz et al., 2014). Stotz et al. (2014) also classify immune response in apoplastic leaf pathogens like *Cladosporum fulvum*, *Zymoseptoria tritici* and *L. maculans* under the broader term of effector triggered defence (ETD). Effectors can be recognised directly and indirectly (Stuart et al., 2013). Dangl and Jones (2001) proposed the 'Guard hypothesis' as a method of indirect recognition which takes into account the fact that Avr proteins may target other host proteins than R proteins (Figure 1.2). They exemplified the working of this model in *Pseudomonas syringae.* The NBS-LRR Prf protein in the host guards Pto, which is a part of host defence, against AvrPto (Dangl and Jones, 2001). Prf initiates the immune response upon recognising the AvrPto-Pto interaction (2001). Another scenario could be when the R protein is activated upon being released from its bound state to the target protein. This occurs when the effector protein interacts with the target protein (Figure 1.2). Other examples that employ the Guard model are *Arabidospsis thaliana* whose RIn4 protein co-exists with RPM1. When *P. syringae* AvrRPM1 degrades RI4, RPM1 is released and activated (Dodds and Rathjen, 2010).

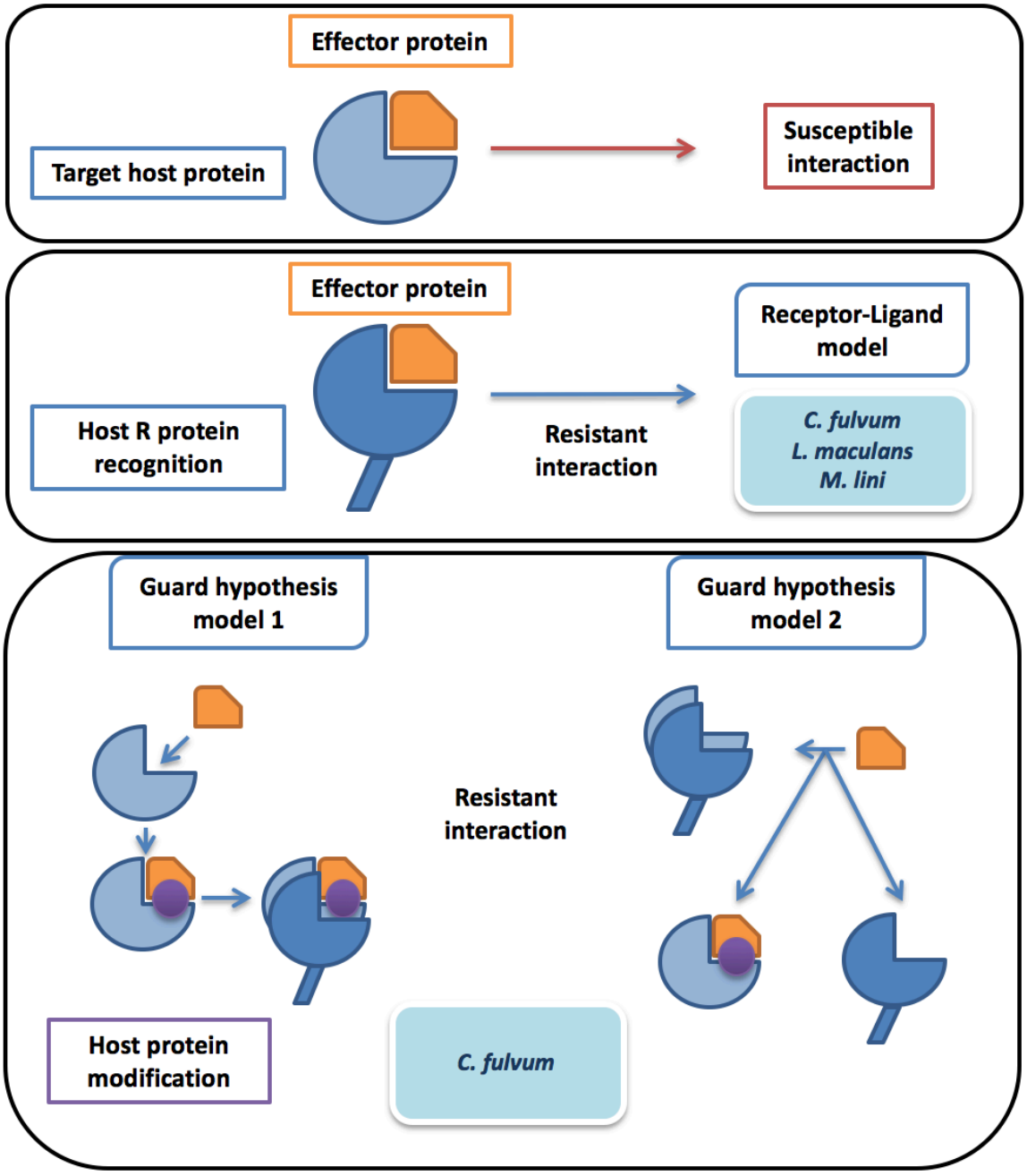

Figure 1.2 Proposed models of interaction between *Avr* and *R* gene products (Susceptible interaction-When the effector protein interacts with its target in the host cell leading to infection; Receptor-ligand model-Where the effector protein is recognised by the host resistance protein leading to a resistant interaction, exemplified in *C. fulvum*, *L. maculans* and *Melampsora lini*; Guard hypothesis model 1-When the effector binds to its target protein which is subsequently modified and this modification is recognised by a R protein leading to immune response, exemplified in *C. fulvum*; Guard model hypothesis 2-When the guarding R protein exists in a bound state with the target protein there is no immune response. Upon effector protein binding to the target, the R protein is released to initiate immune response [Adapted from Stuart et al. (2013)]

Flor based his hypothesis through observations made in the flax-flax rust model system (Flor, 1942). Flax (*Linum usitatissimum*) *R* genes coding for TIR-NBS-LRR proteins interact inside the plant cell with flax rust (*Melampsora lini*) *Avr* genes (*AvrL567, AvrM, AvrP123,* and *AvrP4*) that are translocated into the cytoplasm (Zander et al., 2015). AvrM protein interacts directly (direct recognition; Figure 1.2) with the M resistant protein, inside the host cell (Ve et al., 2013).

Another apoplastic leaf pathogen, *C. fulvum* causes leaf mould on tomato (*Solanum lycopersicum*). *Avr2*, *Avr4*, *Avr4E*, *Avr9* and Ecp1 (Extracellular protein-encoding), Ecp2, Ecp4, Ecp5, Ecp6 and Ecp7 of *C. fulvum* interact with tomato *Cf* genes that fall under the class of Receptor-Like Proteins (RLPs) (Ökmen and de Wit, 2012). *C. fulvum* effectors are secreted into the plant apoplast where *Avr2* behaves offensively and *Avr4* and Ecp6 behave defensively (Ökmen and de Wit, 2012). *Avr2* inhibits plant cysteine proteases that increase the susceptibility of the host to infection (Ökmen and de Wit, 2012). *Avr4* functions to protect the pathogen against tomato chitinases wherein it prevents hydrolysis of chitin by binding to it (Stotz et al., 2014). Studies based on the interaction between *Avr2-Cf-2* indicate that it closely resembles the guard hypothesis (Ökmen and de Wit, 2012). *Avr2* targets Pi-p1 and Rcr3pim which are tomato proteases but it is checked by *Cf-2*, which mounts a hypersensitive response upon structural modification of these proteases by *Avr2* activity (Ökmen and de Wit, 2012) (Figure 1.2). On the other hand, *Avr4-Cf-4* interaction is reported to be direct, like the interaction observed by Flor in flax rust-rust (Dodds and Rathjen, 2010; Ökmen and de Wit, 2012) (Figure 1.2).

L. maculans interacts in a gene-for-gene manner with its host *Brassica* plants (Ansan-Melayah et al., 1998). This means that specific *Brassica Rlm* (Resistance to *L. maculans*) genes can recognise specific avirulence gene (*AvrLm*) products of the pathogen (eg *Rlm6* in *B. juncea* recognises *AvrLm6* in *L. maculans*) (Direct recognition; Figure 1.2). In the presence of both these genes, disease resistance mechanisms are activated in the plant upon infection (Dangl and Jones, 2001) and the pathogen cannot cause disease. However, if the pathogen overcomes host resistance and harbours a second avirulence gene, infection can still occur. This feature of the plant-fungal pathosystem has been used to design disease-resistant crops (Hayward et al., 2012). Recently, the protein structure of *AvrLm4-7*, an avirulence effector of blackleg, has been characterised (Blondeau et al., 2015). Although they did not deduce the exact interaction of *AvrLm4-7* with *Rlm4* and *Rlm7*, they concluded that the interaction could be direct or indirect based on the involvement of different residues in the avirulence protein (Blondeau et al., 2015).

1.4.1 Pathogenicity genes

Pathogenicity genes enable fungi to cause infection in plants. Reduced pathogenicity in some strains may be a function of misregulation or loss of these pathogenicity genes. In recent times, the definition of the term 'effector' is being used to encompass molecules who participate in the infection process but whose exact definition is unknown (Hogenhout et al., 2009). A variety of effector genes or effectors participate in the process of causing infection, which involves the secretion of different compounds into numerous host cell parts (Hogenhout et al., 2009). These may comprise apoplastic or cytoplasmic effectors, delivered into the extracellular space or directly into the plant cell respectively (Sperschneider et al., 2015). Effectors also include genes that perform a wide variety of functions such as avirulence effectors, effectors that play a role in toxicity, enzymes responsible for cell wall degradation, effectors triggering host responses or PAMPs, to name a few (Hogenhout et al., 2009; Sperschneider et al., 2015). In blackleg, other pathogenicity effector genes have been described. These include 3-Ketoacyl-CoA thiolase encoded by the THIOL gene (Elliott and Howlett, 2006), and the *Lmgpi15* gene (Remy et al., 2008b), *LmIFRD* gene (Van de Wouw et al., 2009), Leloir pathway enzyme encoding the *Lmepi* gene (Remy et al., 2009) and *Ipa* gene (Elliott et al., 2008), the *Lmpma1* gene that encodes the plasma membrane isoform of H^+ATPase (Remy et al., 2008a) and the *LmSNF1* (Feng et al., 2014) gene that plays a role in degrading plant cell wall, spore attachment and germination.

1.4.1.1 Avirulence effector genes

As defined by Oku, effector genes usually possess three distinct traits- the abilities to enter the host plant and cause infection after overcoming the host resistance response (Ichinose et al., 2013). Avirulence (*Avr*) effector genes are phytopathogenic genes encoding a small-secreted protein that facilitate in causing infection on host plants or play a role in inducing immune responses upon infection (Rouxel and Balesdent, 2005; Rouxel et al., 2011). The products of these genes are known as effector proteins, and these carry out functions such as plant defence suppression and host cell-wall modifications upon infection (Rouxel et al., 2011). Of the eleven *L. maculans Avr* genes identified, six (*AvrLm1*, *AvrLm6, AvrLm4-7, AvrLm11, AvrLm2* and *AvrLm3*) (Fudal et al., 2007; Gout et al., 2006b; Parlange et al., 2009; Balesdent et al., 2013; Ghanbarnia et al., 2014; Plissonneau et al., 2015), have been characterised (Table 1.1). A seventh *AvrLm* gene, *AvrLmJ1*, virulent on *Brassica juncea* cultivars, has also been characterised (Van de Wouw et al., 2014). Of these 11 genes, seven genes were thought to be located within two clusters in the blackleg genome; *AvrLm1-2-6* and *AvrLm3-4-7-9* (Rouxel and Balesdent, 2005; Van de Wouw et al., 2010). These avirulence genes display features typical of effector genes in that they are expressed early in

infection and secrete small proteins into the apoplast (Stergiopoulos and de Wit, 2009). Identifying novel effectors in fungi is a challenging process due to low sequence similarity to effectors within the same species or distantly related species and no structural homology (Ellis et al., 2009; Sperschneider et al., 2015).

Table 1.1 Currently known avirulence genes of *L. maculans* and their method to overcome host detection

***AvrLm* gene**	**Method of conferring virulence**	**Reference**
AvrLm1	Deletion	(Gout et al., 2006b)
AvrLm2	Single point mutation	(Ghanbarnia et al., 2014)
AvrLm3	Masked by functional *AvrLm4-7*	(Plissonneau et al., 2015)
AvrLm4-7	RIP mutation and single point mutation	(Parlange et al., 2009)
AvrLm6	Deletion, RIP mutation	(Fudal et al., 2007)
AvrLm11	Loss of minichromosome	(Balesdent et al., 2013)
AvrLmJ1	Single point mutation	(Van de Wouw et al., 2014)

Rouxel et al. (2011) mapped *AvrLm1* and *AvrLm6* to SC_6 and *AvrLm4-7* to SC_12 in the *L. maculans* genome, respectively. The *AvrLm1* and *AvrLm6* genes were positioned within a RIP affected *Avr1-2-6* region of the genome that contains two kinds of isochores. Isochores can be defined as long stretches of DNA homogenous in GC content, with abrupt changes in GC content from one segment to another (Parlange et al., 2009). One of the isochores is composed of truncated (degenerate) or complete retrotransposons with a low GC content of 30-40%. The second has a 52% GC content and is gene-rich but devoid of retrotransposons. The *AvrLm2* gene was also found on SC_6 in a cluster with *AvrLm1* and *AvrLm6*. It is a 699 bp gene that encodes a 232 aa protein (Ghanbarnia et al., 2014), located on a GC island in the middle of an AT isochore. Similar to other known avirulence genes, this gene too encodes a small secreted protein and does not match other proteins when searched in databases. The product was found to be secreted 3 days post inoculation (3 dpi) and its expression peaked at 5 dpi, decreasing after that period. Isolates containing *AvrLm2* overcome *Rlm2*-mediated recognition by two point mutations, both which give rise to an amino acid change at Gly^{133} (G^{397} to A/C^{397} and G^{398} to A^{398}).

The region surrounding *AvrLm1* is 296 kb long, rich in GC-isochores and long-terminal repeat retrotransposons and gene poor (Gout et al., 2006b). The AvrLm1 protein itself contains a single cysteine residue and is composed of 205 amino acids. The region comprising *AvrLm6* is 133 kb long, non-coding and contains long-terminal repeat retrotransposons (Fudal et al., 2007). Both genes generate small protein products with a single N-terminal signal peptide. *AvrLm1* and *AvrLm6* have both been found to play a role during infection. However, in races virulent on *B. napus* cultivars that contain the resistance gene *Rlm1* (Resistance to *L. maculans* 1), *AvrLm1* has been noticeably absent (Fudal et al., 2009). This led to the conclusion that the pathogen adapts to *Rlm1* resistance via deletion of the *AvrLm1* locus.

The *AvrLm4-7* locus shows characteristics similar to that of *AvrLm1* and *AvrLm6*. The gene encodes a cysteine-rich protein composed of 143 amino acids, which is secreted early during infection (Parlange et al., 2009). The *Avrm4-7* locus also undergoes deletion to overcome host resistance conferred by *Rlm4* and *Rlm7*. Those strains that infect *Rlm4*-containing plants have been found to undergo a point mutation causing an Arg to Gly substitution. The fitness level of these fungal strains is lower compared to the wild-type strains, but the mutation does not affect *AvrLm7* specificity (Parlange et al., 2009). In this manner, the pathogen evades *Rlm4*-mediated recognition. This region is also known to be affected by RIP mutations. Recently, it was found that the SNP responsible for the mutation in the *AvrLm4* gene making those strains virulent to its corresponding resistance gene, *Rlm4* was characterised using pyrosequencing to assess the frequency of the virulence allele in the sample pool (Van de Wouw and Howlett, 2012). The assay was validated by Sanger sequencing of the entire gene. Mutations at any one of three detected polymorphic sites enable the pathogen to overcome *Rlm7* recognition (Blondeau et al., 2015).

AvrLm11 is the avirulence gene said to confer virulence towards *B. rapa* cultivars that contain the *Rlm11* resistance gene and is hosted on the CDC, SC_22 (Balesdent et al., 2013). As mentioned earlier, this gene too shares common features including coding a cysteine-rich protein, located in the middle of a large 321 kb AT-block and highly expressed 7 dpi, with other avirulence genes. Its location on the dispensable chromosome is what makes this avirulence gene different. The authors concluded that the minichromosome is lost during the sexual cycle in 4.8% of instances and is maintained in the genome over a ten-year interval, amongst field isolates. This meant that maintaining the CDC conferred a selective advantage to the pathogen, as was observed by the reduced fitness of the progeny which did lost the CDC. The complete loss of the CDC was the method to confer virulence towards the matching resistance gene.

AvrLm3 is the most recently discovered avirulence effector of blackleg. It was found to be located in the previously reported cluster of *Avr3-4-7-9* in a telomeric region on SC_12 (Plissonneau et al., 2015). *AvrLm3* recognition by *Rlm3* was found to be 'hidden' by the presence of a functioning allele of *AvrLm4-7*. In the absence of a functional allele, *AvrLm3* is detected again. It was also proposed that this effector may be important to the fitness of the pathogen.

Lastly, *AvrLmJ1* is a recently characterised avirulence gene that confers virulence on *B. juncea* cultivars, which is located in a gene poor region and is cysteine rich. However, it overcomes host recognition via a single point mutation that causes a premature stop codon leading to a truncated protein (Van de Wouw et al., 2014).

Grandaubert et al. (2014b) have proposed that fungi maintain genomes riddled with TEs in order to easily adapt to host plant pressures. They host effectors in dynamic plastic regions of the genome that affords them the capacity to easily evolve leading to rapid sequence diversification, when faced with a novel host. Moreover, it also enables the pathogen to better adapt to existing hosts, a process that can be employed by the pathogen over and over again in the long-term. By hosting effector genes in genomic regions containing repeat elements, the leaking effects of RIP, induced and originally acting on TEs, acts on 'compromised' effector genes leading to their diversification (Rouxel et al., 2011). Also, mutations generated by RIP initiate diversifying selection of effector genes that need to be maintained in the genome and prevent them from going extinct. On the other hand, effectors damaging to pathogen fitness can just as easily be lost. Four of the known effector genes, *AvrLm1*, *AvrLm6*, *AvrLm4-7* and *AvrLm11* are hosted in AT-rich isochores of the blackleg genome (Gout et al., 2006b; Fudal et al., 2007; Parlange et al., 2009; Balesdent et al., 2013; Grandaubert et al., 2014b).

Based on combined knowledge gleaned from the discovery of these avirulence genes, some common features are summarised as follows:- (1) They encode small secreted proteins comprising 200-300 aa (2) Cysteine rich (3) Usually located in an AT-rich region, devoid of other genes (4) Share little to no similarities to known effectors within species or with other species (5) Gene expression is highly upregulated 7 and/or 14 dpi when inoculated on susceptible cultivars (6) The pathogen overcomes host recognition by employing a variety of strategies at the avirulence gene loci like deletion, point mutation leading to amino acid changes or causing stop codons or RIP mutations.

1.5 Whole genome sequencing

SNP discovery from next generation sequencing data is now possible, with advances in Illumina genome sequencing technologies providing over 40 gb of sequence data in a single sequencing lane (Edwards et al., 2013). Genetic variation between individuals can be identified using re-sequencing, which can provide molecular genetic markers and insights into gene function (Imelfort et al., 2009). During whole-genome re-sequencing of different samples, millions of reads are aligned to a reference genome sequence. This then enables the determination of nucleotide variants between the samples.

A reference genome is a valuable resource for future applications and experiments, including marker discovery, re-sequencing of other samples and studies of genome diversity. A reference genome has proven to be useful in other plant pathogens such as *Fusarium graminearum* (Cuomo et al., 2007) and *Stagonospora nodorum* (Hane et al., 2007). The availability of a reference genome can help in conducting other analyses such as transcriptome construction, RNA seq analysis and annotations on a reference genome help decipher the genetic content and its biological meaning (Kuo et al., 2014). A fully-annotated reference genome is extremely useful in performing metabolomics and identifying secondary metabolite genes (Kuo et al., 2014). Furthermore, a reference genome can be used to identify variations in the sequence and other mutants as the first step into population genetics. Uses also extend to phylogenomics, ecogenomics and comparative genomics (Kuo et al., 2014). The small genome size of the pathogen makes it cost-effective and easy to re-sequence other population samples.

1.6 Single Nucleotide Polymorphisms (SNPs)

A single nucleotide difference between two individuals at a particular location can be classified as a single nucleotide polymorphism (SNP). Currently, SNPs are the most popular marker of choice for fine mapping of heritable traits (Chagné et al., 2007). SNPs can be categorised as transversions (C/G, A/T, C/A or T/G), transitions (G/A or C/T) and insertions/deletions (indels). Studying the genetic diversity of fungal genomes using SNPs as genetic markers is becoming an increasingly popular approach due to several beneficial features (Appleby et al., 2009). SNPs are found in high densities in genomes, which assists in applications like map-based positional cloning, creating high-density genetic maps and genome mapping (Edwards and Batley, 2004; Edwards et al., 2007; Duran et al., 2010). Furthermore, SNPs are also a very good indicator of genetic diversity within fungal pathogenic species such as *L. maculans.* They are stable during evolution, are not affected by

selection pressure and have low mutation rates which allow for an easy analysis of complex genetic traits and difference in levels of diversity (Appleby et al., 2009)

1.6.1 Marker assisted studies

Genomic diversity and evolution studies using SNPs are assisted by large-scale genotyping assays like the Illumina GoldenGate assay, which are cost-effective for analysing a large number of SNPs in multiple individuals. The Illumina GoldenGate assay is used to analyse 384 to 3072 SNPs in 96 individuals simultaneously. SNP genotypes are differentiated using allele-specific and locus-specific oligonucleotides (ASO and LSO). Studies have shown that SNPs are highly informative when available in large numbers (Liu et al., 2005) and that they are less error-prone than microsatellites, as seen in zebra-finch studies (Ball et al., 2010). To date, no large-scale diversity studies using SNPs have been conducted in *L. maculans*.

SNP discovery has also been conducted in other fungal species such as *Leptographium longiclavatum* (Ojeda et al., 2014) and *Venturia inaequalis* (Celton et al., 2010). The former developed a SNP marker resource to analyse the population of *L. longiclavatum* in Canada and USA (Ojeda et al., 2014). They were also used to identify candidate adapter SNPs that play a role in pathogenicity and host defence detoxification (Ojeda et al., 2014). The latter used a bin mapping approach to identify SNPs via whole or partial genome resequencing of individuals in the progeny (Celton et al., 2010). They used SNPs as anchors in the draft genome in order to assemble and order genome contigs (Celton et al., 2010). Tollenaere et al. (2012) used transcriptome data for the identification of 27 SNP loci which they used to analysed multilocus genotypes across 380 samples of *Podosphaera plantaginis*. They concluded that the genetic diversity of this pathogen was low and SNP markers were an ideal tool to analyse this as compared to microsatellite markers.

In blackleg, Zander et al. (2013) used SGSautoSNP (Lorenc et al., 2012) to identify over 21,000 SNPs. The frequency of SNPs in the genome was one SNP per 2065 bp. They found low SNP density and subsequent sequence diversity in regions such as SC_30 which comprises mitochondrial DNA and other regions consisting of housekeeping genes. Overall, the predicted SNPs were highly conserved and polymorphic amongst the Australian *L. maculans* isolates. They propose the use of their novel SNP resource to study the genetic diversity of this pathogen in Australia.

1.7 Gene loss

Gene loss is a major process employed by eukaryotes to evolve, especially under selection pressure (Krylov et al., 2003). Loss of genes at this scale is also evident in other pathogens such as *Encephalitozoon cuniculi* (Katinka et al., 2001). Pseudogenisation and gene loss have also been reported in humans, since the lineage diverged from chimpanzees (Wang et al., 2006). Gene loss is not limited to eukaryotes and has also been analysed in prokaryotes (Krylov et al., 2003). For example, genomes of bacteria such as *Buchnera aphidicola* have undergone major gene loss and lost up to 86% of their genes in order to adapt to the endosymbiotic lifestyle (Baumann et al., 1995). Gene loss and horizontal gene transfer (HGT) account for the two most significant processes that bring about large-scale changes in the genomes of these organisms (Krylov et al., 2003).

The genomes of filamentous plant pathogens are shaped by several processes, gene loss being one of them (Raffaele and Kamoun, 2012). There is also considerable variation in genomic content between species like *Ustilago maydis* that contains almost 1000 fewer genes than *Blumeria graminis*, despite having a larger genome (Raffaele and Kamoun, 2012). Spanu et al. (2010) found that the powdery mildew genome (*B. graminis* f.sp *hordei*) is affected by gene loss, effects of retrotransposons and genome expansion. The genes lost by the pathogen are secondary metabolites or carbohydrate-active enzymes. The authors report that these genes are lost due to their redundancy in the current lifestyle of the pathogen. *L. maculans* is a prime example where gene loss and pseudogenisation occur in virulent strains of the pathogen (Raffaele and Kamoun, 2012). All isolates virulent on *Rlm1* and half of the number virulent on *Rlm6* containing cultivars display deletions at the *AvrLm1* and *AvrLm6* loci respectively (Fudal et al., 2009; Van de Wouw et al., 2010). Retrotransposons found in AT-rich regions of the genome also play a role in gene deletion and loss (Raffaele and Kamoun, 2012). Virulent strains of *M. oryzae* also display deletion or pseudogenisation of *Avr-Pita* (Dai et al., 2010; Chuma et al., 2011). Gene loss is an important phenomenon occurring in plant pathogens that not only contributes to enabling virulence but also in rapid evolution against the host.

1.7.1 Presence/Absence Variation (PAVs)

Presence/Absence variation is defined as "sequences that are present in one genome, but entirely missing in another genome" (Wang et al., 2014). There have been reports of PAVs in plants (Batley et al., 2003; Zhang et al., 2014), humans (Iafrate et al., 2004), bacteria (Arrach et al., 2008) and fungi (Huang et al., 2014). Genes involved in pathogenicity and other stress-response genes were found to be present/absent in *Cucumis melo* and *A. thaliana* (Shen et al., 2006; Gonzalez et al., 2013). McHale et al. (2012) reported the connection of PAVs to processes such as biotic response and nucleotide binding. Wang et al. (2014) found 28,912 PAV markers amongst 33 whole-genomes of soybean which they used to create a powerful marker resource along with SSRs. PAVs were also purported to be the reason behind *M. oryzae* evading the host immune response (Huang et al., 2014). Zander (2015) details the detection of PAVs in the *L. maculans* genome which highlighted selection pressure in some regions of the genome and successfully validated the presence/absence of avirulence genes like *AvrLm1* and *AvrLm6*.

1.8 Research aims

This thesis aimed to contribute to the pool of knowledge about blackleg, its genomics and population diversity in Australia. Specifically, we sought information about the genomic structure of the pathogen, to identify novel avirulence effectors and to identify the genomic diversity of this pathogen on a national level. Whole-genome sequence data was used to gain information at the genomic level about the structure and function of this pathogen.

Chapter 3

We aimed to analyse the population diversity of this pathogen in Australia using SNPs as molecular markers. We hoped to gain an overall picture of this pathogen by sampling from different canola growing regions across the country. The primary aim was to identify significant association of SNP haplotype blocks to parameters such as state or location of isolation of sample, year, stubble cultivar isolated from or stubble resistance source. These haplotype blocks could also identify regions of selection pressure in the genome that could be correlated to what cultivar was grown in the field and subsequently identify novel avirulence genes in the genome.

Chapter 4

Based on axiomatic knowledge, avirulence genes are deleted from the genome to confer virulence on host plants. Therefore, we aimed to conduct gene loss analysis in isolates of *L. maculans* to glean what is lost by the pathogen. We hypothesised that the results of this analysis would highlight candidate avirulence effectors and also genes important to the fitness of the pathogen.

Chapter 5

Previously established PAV and SNP resources were combined with the gene loss data resource to lay the foundation of a pan-genome and decipher black genomics. We primarily aimed to identify avirulence effectors that could be deleted, highly RIP mutated, affected by SNPs or highlighted as highly variable in the PAV data. We also wanted to exemplify how this mega-data resource could be used to further characterise the blackleg genome.

Disease mediation, recognition and regulation all occur through different effectors in *L. maculans*. Therefore, the aim of identifying these effectors was a common theme visible across all chapters of this thesis. Utilising bioinformatics tools also helps establish other genomic characteristics and population information that is extremely useful for future studies.

2 Materials and Methods

2.1 Sample growth and storage

2.1.1 Media preparation

Solid media

Solid media was prepared using Campbell's V8 juice (Campbells Australia Pty Ltd). V8 juice was first strained through a miracloth to separate the liquid component of the juice from the fibrous component. The strained V8 juice was then stored in 40 ml aliquots in falcon tubes or Schott bottles at -20°C.

10% V8 Agar media was made as follows- 40 ml of strained V8 juice was combined with 1.5 g calcium carbonate ($CaCO_3$) and 8 g of Agar. 20% V8 Agar was made by adding 80 ml of strained V8 juice. This mixture was made up to 400 ml with Milli-Q water (Merck Millipore Corporation) in a 1 L Schott bottle, ensuring that the pH was adjusted to 6.5. The media was then autoclaved at 120°C for 20 minutes and allowed to cool down to a temperature of ~45-50°C. Once cooled, 10 mg/ml MeOH of Rifampicin and 30 mg/ml EtOH of Chloramphenicol antibiotics were added to the mixture. Then, operating in a laminar flow cabinet, the media was poured into 90 x 14 mm petri dishes for use. Once the media was dry, the plates were covered, sealed with clear wrap and stored at -2°C.

Liquid media

10% V8 liquid media was prepared by making up 40 ml of strained V8 juice to 400 ml with Milli-Q water (Merck Millipore Corporation) (pH adjusted to 6.5) and autoclaving at 120°C for 20 minutes. Once the media was cooled to ~45-50°C, antibiotics were added; 10 mg/ml MeOH of Rifampicin and 30 mg/ml EtOH of Chloramphenicol. Operating in a laminar flow, the liquid media was then poured into tissue culture plates measuring 145 x 20 mm, for use.

2.1.2 Sample growth

Solid media growth

Fungal isolates were grown on 10% V8 Agar media. The isolates were grown in 9.5 cm plates sealed with Parafilm M® (Bemic Company Inc.) in a growth cabinet at room temperature in a 12-hour dark (dark light by Crompton 18 Watt F18T8 BLB bulb) and 12-hour light cycle (for increased spore production) for two weeks. 6 mm Grade AA Whatman™ filter discs (GE Healthcare Life Sciences), are placed at 1 inch intervals to obtain a source for future fungal growth.

Liquid media growth

Fungal isolates were grown on 20% V8 juice liquid media to obtain better DNA yield from tissue after extraction. 50 ml of liquid media was poured into 9.5 cm plates which were then inoculated with a filter disc or plug. This inoculum was obtained from previous plates grown in the lab, a 0.5 cm square of agar. The conditions of growth detailed for solid media growth were also used for liquid media growth.

2.1.3 Storage

The isolates were either stored in liquid form (agar piece in water) or filter form (Grade AA, Whatman™ (GE Healthcare Life Sciences) filter discs in Parafilm M®-sealed 1.5 ml Eppendorf tubes in silica beads). Fungal isolate sources were replenished every year to ensure viability.

2.2 Spore growth, extraction and storage

Solutions of spores from the isolates were used to perform inoculation on cultivars for infection studies. First, fungal isolates were grown on solid media as detailed above in Section 2.1.1. After three weeks, 10 ml of autoclaved Milli-Q water (Merck Millipore Corporation) was used to flood the plates and the surface was agitated using a spreader. A sterile funnel was lined with an autoclaved piece of Miracloth (Merck Millipore Corporation) and the contents of the plate were drained through it into a sterile falcon tube. Spore solutions were stored at -20°C within one hour of harvest.

To quantify spores in the solution, ~10 µl of spore solution was pipetted onto a hemocytometer (Neubauer). A light microscope (Olympus) was used at x100 magnification to count spores. If the spores visible were too many to count, the solution was diluted 1 in a 100 or 1 in a 1000 using autoclaved Milli-Q water (Merck Millipore Corporation) and the process is repeated.

2.3 Tissue extraction

After three weeks of regrowing the isolates on liquid media, fungal tissue was harvested from the plates. An autoclaved piece of Miracloth (Merck Millipore Corporation) was placed in a funnel and the plate contents were drained. A scraper was used to ensure that all tissue was collected from the plate. Excess liquid from the cloth was removed by squeezing the leftover fungal tissue. Sterilised tweezers were used to collect the tissue in the cloth and placed in falcon tubes which were then stored at -80 °C for later DNA extractions.

2.4 DNA extraction

2.4.1 Microprep DNA extraction

Buffer preparation

Three individual buffers are required for the microprep protocol for DNA extraction-
Nuclei lysis buffer- Composed of 0.05 M EDTA, 2 M NaCl, 2% CTAB and 0.2 M Tris
DNA extraction buffer- Composed of 5 mM EDTA, 0.35 M sorbitol and 0.1 M tris-base
Sarkosyl-Solution made up to 5% w/v concentration
(These buffers are stored at room temperatures and can be used for up to six months)

These three buffers constitute the final microprep buffer. In order to make the microprep buffer, 2.5 parts of nuclei lysis buffer was mixed with 2.5 parts of DNA extraction buffer and 1 part of 5% Sarkosyl. Next, 0.3-0.5 g of sodium bisulphite was added per 100 ml of microprep buffer.
(Microprep buffer must only be made for immediate use and cannot be stored)

Microprep DNA extraction protocol

The microprep protocol for DNA extraction described by Fulton et al. (1995) was followed for fungal genomic DNA extraction, with the following modifications. Post incubation at 65 °C, 750 µl of Phenol: Chloroform was used for washing each sample. This was followed by centrifugation for 5 minutes at 10,000 rpm. 50 µl of DNase-RNase free water was used to resuspend DNA samples. Reactions sensitive to phenol contamination were washed instead with Chloroform: Isoamyl.

2.4.2 96 Plant DNeasy kit

In order to obtain high quality DNA free of secondary metabolite contamination that usually occurred in DNA obtained from the microprep protocol, kit extraction method was used. These contaminants are known to interfere with library preparation reagents.

The DNeasy® Plant Mini Kit (Qiagen) was used according to manufacturer's instructions. The only modification performed was the grinding of fungal tissue with a mortar and pestle prior to adding to the QIAschredder column.

2.5 Quality assurance and DNA quantification

Quality assurance and quantification of each sample was conducted post DNA extraction to know the quality and quantity of DNA content before the next step of next generation sequencing library preparation.

2.5.1 Quality assurance- Gel electrophoresis

DNA samples were run on a 1% TAE-Agarose gel to test the quality of the DNA. A 1%-TAE-Agarose gel was prepared by combining 1% agarose with 1xTAE buffer and 0.5 µg/ml of ethidium bromide. The mixture was heated in a conical flask in the microwave and swirled at 1-1:30 minute intervals to homogenously combine the agarose with the TAE buffer. Once the mixture was combined, it was allowed to cool and ethidium bromide was added. Next it was poured into a gel tray with well-combs (type of comb used depends on the number of samples being analysed). Once the gel was set, it was removed from the tray and transferred into a gel tank containing enough 1xTAE buffer to sufficiently cover the surface of the gel when placed in the tray. 1-2 µl of each DNA sample was combined with 1X concentration of 6X loading dye (ThermoFisher Scientific). One combination of sample and dye was loaded per well of gel. 6 µl of 1 kb GeneRuler™ ladder (0.5 µg/ µl, 50 µg; ThermoFisher Scientific) was added in a separate well to gain a preliminary concentration of the DNA sample. The gel was run at 80-100 volts for 20-60 minutes depending on the size of the gel.

Once the run was completed the gel was transferred to the Gel documentation system for imaging. Once the chamber was closed, UV light was turned on to visualise the fluorescent DNA, caused by the binding of ethidium bromide to the DNA. In order to gain an adequate image, the Major Science™ SmartView Pro 1200 Image System (Major Science) was used.

2.5.2 Quantification- Qubit

After DNA extraction, each sample was quantified using a Qubit fluorometer (Life Technologies) according to the manufacturer's instructions. In order to quantify the samples whose DNA concentration was less than 10 ng/µl, high sensitivity (HS) reagents were used. For samples whose concentration was between 10-500 ng/µl, broad range (BR) reagents were used.

2.6 Whole genome sequencing

Whole genome sequencing was performed using the Illumina GAIIx, Illumina Hiseq and Illumina Miseq platforms. The first two platforms were hosted at the Australian Genome Research Facility (AGRF) and the latter, in the Batley laboratory. Libraries required for sequencing were made as per the manufacturer's instructions (Illumina Inc).

2.7 Data analysis

R program for statistical computing (R Core Team) was used to perform statistical analyses in this thesis. Geneious (Version 6.1.8; Kearse et al., 2012) was used to perform genomic sequence analyses.

3 Population diversity of *Leptosphaeria maculans* in Australia

Published in International Journal of Biology (2015), Vol 7, No 3. Accessed at- http://dx.doi.org/10.5539/ijb.v7n3p18

3.1 Introduction

The ubiquitous fungal pathogen *L. maculans* is the causal agent of phoma stem canker (blackleg) in *Brassica napus, B. juncea, B. rapa* and *B. oleracea*: canola, vegetable and mustard crops (West et al., 2001). Annually, this pathogen causes an average loss of AUD $76.6 million to the Australian economy (Murray and Brennan, 2012). Furthermore, *L. maculans* isolates present in Australia are classified as highly virulent, able to cause disease even in the more resistant Brassica species: *B. juncea*, *B. nigra* and *B. carinata* (Purwantara et al., 1998).

Daverdin et al. (2012) describe that rapid evolution in pathogens gives rise to new strains to combat crop defences. Maintenance of crop resistance directly depends on the field population size of the pathogen, its evolutionary potential and cropping practices that directly affect its reproductive system (Daverdin et al., 2012). *L. maculans* can survive as a saprobe in the stubble of infected plants for many years and this is usually favoured by dry hot summers and cold winters (West et al., 2001). During this period, it produces sexual inoculum (ascospores), which can travel from several hundred metres to several hundred kilometres (Travadon et al., 2011) and infect plants followed by asexual spore (conidia) production at the site of infection (Rouxel and Balesdent, 2005).

Recent genome sequencing revealed that the *L. maculans* genome has an isochore-like structure (Rouxel et al., 2011), where the genome is divided into AT and GC-rich blocks, probably caused by the amplification of transposable elements and repeat-induced point (RIP) mutations. The RIP mechanism causes nucleotide substitutions from C to T and G to A and is a premeiotic repeat-inactivation mechanism specific to fungi that creates genetic diversification in the fungal genome (Rouxel et al., 2011). Fudal et al. (2009) reported that RIP affects the *AvrLm6* locus, causing gene inactivation and leading to virulence. Furthermore, isolates have been found to undergo continuous deletions and mutations apart from RIP that lead to further genome diversity. Current approaches to establish blackleg resistance in canola have not been successful in fully controlling this pathogen (Hayward et al., 2012). Therefore, understanding how this fungus has evolved, diversified and

spread in Australia is important in providing information for the breeding and sowing of improved resistant varieties of *Brassica*.

Population diversity in *L. maculans* has been examined using a variety of markers. Genetic differences between eastern and western Australian isolates have previously been found using microsatellites and minisatellites, which were attributed to the presence of arid desert between the coasts (Hayden et al., 2007). Minisatellite markers used to analyse four field populations in France found high levels of gene and genotypic diversity within populations and high gene flow between populations, consistent with randomly mating populations (Gout et al., 2006a). Dilmaghani et al. (2012) also used minisatellite markers to show that the *L. maculans* population in Western Canada comprises two genetically distinct populations. A further study implementing fourteen minisatellite markers also found clonal sub populations of this pathogen on *B. oleracea* in Mexico (Dilmaghani et al., 2013). However, Travadon et al. (2011) found the French *L. maculans* population to be panmictic. This study also employed minisatellite and microsatellite markers. Other investigations into blackleg population structure have also been conducted using amplified fragment length polymorphisms (AFLPs) (Purwantara et al., 2000) and restricted fragment length polymorphisms (RFLPs) (Barrins et al., 2004).

SNPs have recently become a popular choice of molecular marker for population diversity studies, and offer significant benefits in terms of abundance in genomes and ease of high-throughput assessment. SNPs are single base-pair differences between two individuals at a particular locus (Appleby et al., 2009). SNPs can be classified as transitions (C to T, G to A), transversions (C to G, A to T, T to G or C to A) and insertions/deletions (indels) of a single base pair. Such molecular markers are good tools to analyse the various processes encompassing the population genetics and evolutionary processes of an organism. These include mating systems, patterns of speciation, dispersal, mutation, migration and selection etc. (Giraud et al., 2008).

The Illumina GoldenGate genotyping assay can be used to simultaneously analyse 384-3072 SNP loci across multiple individuals (Tindall et al., 2010). Previous studies using the Illumina GoldenGate assay have shown that it can be used to reliably score SNPs for genetic analysis (Durstewitz et al., 2010). Furthermore, it is cost-effective and flexible for analysing large numbers of SNPs (Appleby et al., 2009). We applied 384 previously developed *L. maculans* SNPs (Zander et al., 2013) in a GoldenGate assay to analyse 59 Australian *L. maculans* population isolates collected from different years, regions and cultivars, assessing the diversity of this pathogen across Australia.

3.2 Materials and Methods

3.2.1 Fungal Samples used in assay

The isolates 04MGPP021 and 06MGPP041 provided by Dr Harsh Raman (New South Wales Department of Industry and Innovation) were used for initial SNP prediction and validation. These were used to assay the Skipton x-Ag-Spectrum mapping population segregating for *Rlm4* resistance (Raman et al., 2012). A total of 130 fungal isolates were received from Prof Barbara Howlett's collection at the University of Melbourne. The isolates were chosen from a list of over 2000 within the laboratory collection. They were carefully selected to cover a wide range of parameters including region of collection, cultivar grown at collection site, year isolated and *Avr* gene complement (where data was available; see Chapter 4, Table 4.5). This list was narrowed down to a total of 59 fungal isolates was analysed using the Illumina GoldenGate assay. 96 samples were run in the assay, including replicates and controls (Table 3.1). Of these, Lm-1 and Lm-2 were chosen to be replicated as they were the parents for SNP prediction in this data set. Depending on availability of quality DNA, other isolates replicated belonged to the differential set of isolates for which *AvrLm* gene complement information was available (See Chapter 4, Table 4.5). Some isolates were also replicated at random (Table 3.1). These samples were replicated where enough quality DNA was available. Some samples were also chosen to be replicated at random. The isolates received were either stored in liquid form (agar piece in water) or filter form (filter discs in silica beads) (See Section 2.1.3, Chapter 2).

Table 3.1 List of *L. maculans* population isolates used in this study

(VIC-Victoria; NSW-New South Wales; WA-Western Australia; SA-South Australia; Isolate v23.1.3 is the result of a series of in vitro crosses between European field isolates (Balesdent et al., 2001); Not all data on these isolates was available, "-" denotes an unknown variable. IBCN numbers represent the IDs of "International Blackleg of Crucifers Network" isolates (Marcroft et al., 2012))

As referred to in text	**Isolate**	**Year Cultured**	**Species isolated from**	**Stubble cultivar**	**Stubble collection site**	**Country/State**	**Replicates**	**Reference**
Ref	Reference (v23.1.3)	Mid-1990		-	-	Europe	N/A	(Rouxel et al., 2011)
Lm-1	04MGPS021 (21)	2004	*B. napus*	AG-Emblem	Eyre Peninsula	SA	2	
Lm-2	06MGPP041 (41)	2006	*B. napus*	Skipton	Lake Bolac	Vic	2	
Lm-3	04MGPP003	2004	*B. napus*	TI1 Pinnacle	Geelong	VIC	N/A	
Lm-4	04MGPP008	2004	*B. napus*	Unknown	Wonwondah	VIC	N/A	
Lm-5	04MGPP016 (D11)	2004	*B. napus*	AG-Emblem	Bordertown	SA	N/A	
Lm-6	04MGPP022	2004	*B. napus*	Grace	Moyhall	SA	N/A	
Lm-7	04MGPP026	2004	*B. napus*	Grace	Moyhall	SA	N/A	

Lm-8	04MGPP035	2004	*B. napus*	TI1 Pinnacle	Geelong	VIC	N/A	
Lm-9	04MGPP041	2004	*B. napus*	Grace	Wonwondah - Pymers	VIC	N/A	
Lm-10	04MGPP043	2004	*B. napus*	Grace	Wonwondah - Pymers	VIC	N/A	
Lm-11	04MGPP045	2004	*B. napus*	Grace	Wonwondah - Pymers	VIC	N/A	
Lm-12	04MGPP046	2004	*B. napus*	TI1 Pinnacle	Laharum	VIC	N/A	
Lm-13	04MGPP049	2004	*B. napus*	TI1 Pinnacle	Laharum	VIC	N/A	
Lm-14	04MGPS006	2004	*B. napus*	Surpass 400	Eyre Peninsula	SA	N/A	
Lm-15	04MGPS016	2004	*B. napus*	Surpass 603CL	Bordertown - Ballinger	SA	N/A	
Lm-16	04MGPS024	2004	*B. napus*	ATR-Beacon	Bordertown - Ivan	SA	N/A	
Lm-17	05MGPP002	2005	*B. napus*	ATR-Beacon	Woseley	SA	N/A	
Lm-18	05MGPP033	2005	*B. napus*	Skipton	Yeelana	SA	N/A	
Lm-19	06MGPP019	2006	*B. napus*	ATR-Beacon	Wagga Wagga	NSW	N/A	
Lm-20	06MGPP025	2006	*B. napus*	ATR-Beacon	Wagga Wagga	NSW	N/A	

Lm-21	06MGPS032	2006	*B. napus*	Surpass 501TT	Keith	SA	N/A	
Lm-22	07VTJH002	2007	*B. juncea*	JC05002	Horsham	Vic	N/A	
Lm-23	07VTJH020	2007	*B. juncea*	JC05007	Horsham	Vic	N/A	
Lm-24	D13	2009	*B. napus*	Hyola50	Cummins	SA	N/A	(Marcroft et al., 2012)
Lm-25	09SMJ087	2009	*B. juncea*	EXCEED OasisCL	Kaniva	VIC	N/A	
Lm-26	10SMJ041	2010	*B. juncea*	EXCEED OasisCL	Tamworth	NSW	N/A	
Lm-27	LM300	2002	*B. napus*	TI1 Pinnacle	Mt Barker	WA	1	
Lm-28	LM580	2003	*B. napus*	ATR-Beacon	Wonwondah	Vic	N/A	
Lm-29	LM592	2003	*B. napus*	TI1 Pinnacle	Mt Barker	WA	N/A	
Lm-30	LM659	2003	*B. napus*	Hyden	Wongan Hills	WA	N/A	
Lm-31	LM661	2003	*B. napus*	Hyden	Wongan Hills	WA	N/A	
Lm-32	IBCN13 (D1)	1991	*B. napus*	Unknown	Mt Barker	WA	1	(Balesdent et al., 2005)
Lm-33	IBCN15 (D2)	1988	*B. napus*	Unknown	Streatham	Vic	3	(Purwantara et al., 2000)

Lm-34	IBCN16 (D3)	1988	*B. napus*	Unknown	Mt Barker	WA	3	(Purwantara et al., 2000)
Lm-35	IBCN17 (D4)	1988	*B. napus*	Unknown	Millicent	SA	1	(Balesdent et al., 2005)
Lm-36	IBCN18 (D5)	1988	*B. napus*	Unknown	Penshurst	Vic	3	(Purwantara et al., 2000)
Lm-37	IBCN75 (D6)	1987	*B. napus*	Unknown	Mt Barker	WA	3	(Purwantara et al., 2000)
Lm-38	IBCN76 (D7)	1987	*B. napus*	Unknown	Mt Barker	WA	3	(Purwantara et al., 2000)
Lm-39	D8 (M)	2005	*B. napus*	Surpass 501TT	Mt Barker	WA	3	(Marcroft et al., 2012)
Lm-40	D9 (M)	2005	*B. napus*	ATR-Beacon	Mt Barker	WA	3	(Marcroft et al., 2012)
Lm-41	PHW1223 (D10)	1987	*B. napus*	Unknown	Mt Barker	WA	3	(Purwantara et al., 2000)
Lm-42	V4	1988	*B. napus*	Unknown	Numurkah	Vic	N/A	(Van de Wouw et al., 2010)

Lm-43	35	1988	*B. napus*	Unknown	Penshurst	Vic	N/A	(Van de Wouw et al., 2010)
Lm-44	80	1988	*B. napus*	Unknown	Millicent	SA	N/A	(Van de Wouw et al., 2010)
Lm-45	89	1988	*B. napus*	Unknown	Millicent	SA	N/A	(Van de Wouw et al., 2010)
Lm-46	535	2003	*B. napus*	TI1 Pinnacle	Lake Bolac	Vic	1	
Lm-47	1245	1988	*B. napus*	Unknown	Galong	NSW	N/A	(Van de Wouw et al., 2010)
Lm-48	04S012 (D12)	2004	*B. napus*	Surpass603CL	Bordertown	SA	1	
Lm-49	04S005	2004	*B. napus*	Surpass400	Eyre Peninsula	SA	1	
Lm-50	04P042	2004	*B. napus*	Grace	Wonwondah	Vic	N/A	
Lm-51	05P032	2005	*B. napus*	Skipton	Yeelanna	SA	N/A	

Lm-52	06P039	2006	*B. napus*	Skipton	Lake Bolac	Vic	1	(Van de Wouw et al., 2010)
Lm-53	06S014	2006	*B. napus*	Surpass 501TT	Bordertown	SA	N/A	
Lm-54	06S012	2006	*B. napus*	ATR-Beacon	Bordertown	SA	1	
Lm-55	06S039	2006	*B. napus*	Hyola60	Lake Bolac	Vic	N/A	(Van de Wouw et al., 2010)
Lm-56	06J085	2006	*B. juncea*	Unknown	Horsham	VIC	1	
Lm-57	06J095	2006	*B. juncea*	Unknown	Horsham	Vic	N/A	
Lm-58	06J112	2006	*B. juncea*	Unknown	Horsham	Vic	N/A	
Lm-59	04MGPP029	2004	*B. napus*	TI1 Pinnacle	Geelong	VIC	N/A	

3.2.2 Fungal sample growth, harvest and sample quantification and quality check

See sections 2.1, 2.3, 2.4 and 2.5 of Chapter 2.

3.2.3 Illumina GoldenGate assay

A total of 384 SNPs were selected for the Illumina GoldenGate assay. These were chosen from the list of 21,814 SNPs described in Zander et al. (2013). The SNP prediction was conducted using isolates Lm-1 and Lm-2 (Zander et al., 2013). Approximately 100 bp of sequence on either side of the SNP was scrutinized to ensure that no degenerate bases or other polymorphisms were present. In total, 40 bp of sequence is required upstream and downstream of each SNP in order to run them on this assay (Appleby et al., 2009). A designability assessment conducted by Illumina scored the 384 SNPs at 0.4 or above, which is deemed a good score for the Illumina GoldenGate assay (Durstewitz et al., 2010). The SNPs were chosen to cover the expanse of supercontigs on which SNP changes were detected (Zander et al., 2013). The 27 supercontigs chosen for SNP selection represented 96.7% of the total SNP changes detected in the *L. maculans* genome (Zander et al., 2013). Sample preparation for the Illumina GoldenGate assay was performed according to the Illumina GoldenGate Genotyping Assay guide (Catalogue No. GT-901-1001). Illumina software "GenomeStudio" (Illumina Inc., 2013) was used to manually cluster the SNPs into the two possible genotype clusters (A and B) for this haploid organism. SNPs that clustered confidently were selected for future data analyses, and monomorphic and non-clustering SNPs were eliminated from further analyses, resulting in 214 high-quality SNPs. A sub-data set of 193 was used for linkage disequilibrium (LD) analysis (SNPs polymorphic only in Lm-1 and Lm-2 were omitted).

3.2.4 Data analysis

The data set of 214 SNPs was arranged according to predicted positions of supercontigs on chromosomes, as outlined in Supplementary Figure S1 of Rouxel et al. (2011). The isolates were then sorted into groups based on different parameters detailed in Table 3.1 and the data set was analysed for blocks of SNPs occurring together.

All isolates were considered to be part of one population for the statistical analyses. The phylogenetic tree, linkage disequilibrium (LD) analysis, heatmap and principle component analysis were generated using the statistical computing program R (R Core Team, 2015). The SNP positions were given 1 and 0 values (for dendrogram and PCA) or 'A/A' and 'B/B' (for LD) for each genotype call and 'NA' was assigned to missing values. A Euclidean distance matrix was generated and used to create a phylogenetic dendrogram that underwent 1000 bootstrap iterations. Population LD was calculated using the R package 'genetics' (Warnes et al., 2012). The heatmap was generated to visualise the LD data using the R package 'LDHeatmap' (Shin et al., 2006). PCA was performed using the R packages 'ade4' (Dray and Dufour, 2007) and 'maptools' (Lewin-Koh et al., 2012). SNPs with possible null and private alleles were checked against the reference (Rouxel et al., 2011) using Geneious (Version 6.1.8; Kearse et al., 2012).

3.3 Results

3.3.1 Illumina GoldenGate results

The results from the GoldenGate assay supported the SNP prediction of Zander et al. (2013). The data generated was sorted for SNPs that had high confidence clusters. 2.6% of SNPs had missing values (NA) for all isolates, 29.9% were monomorphic and 11.7% were non-clustering SNPs. SNPs belonging to these three categories were eliminated from further analyses. No correlation between eliminated SNPs and SNP score or supercontig on which they were positioned could be observed. Filtering for quality polymorphic data resulted in a dataset of 214 SNPs. A subset of 193 SNPs was used for LD after elimination of a further 21 SNPs. Reproducibility less than 100% was due to missing data in one or other of the replicates (Table 3.2).

Table 3.2 Percent reproducibility of replicates used in the assay

Isolate name	Replicates	Reproducibility (%)
Lm-1	2	99.22
Lm-2	2	98.96
Lm-27	1	98.96
Lm-32	1	99.74
Lm-33	3	97.92
Lm-34	3	92.52
Lm-35	1	99.22
Lm-36	3	97.20
Lm-37	3	96.26
Lm-38	3	96.26
Lm-39	3	93.93
Lm-40	3	92.06
Lm-41	3	94.39
Lm-46	1	95.57
Lm-48	1	98.44
Lm-49	1	98.96
Lm-52	1	97.40
Lm-54	1	98.96
Lm-56	1	99.22

3.3.2 General Marker and Population Statistics

The SNP data was mined for possible private alleles (Table 3.3). Private alleles are unique alleles in isolates that denote genetic distinctiveness. Of the 19 private allele SNPs, seven occurred in the isolate Lm-1 and six occurred in the isolate Lm-2, both of which were used for SNP discovery. Of the other four alleles occurring at low frequency, four were in Lm-2 and Lm-55 only, and two were present in only Lm-2 and Lm-24 and Lm-2 and Lm-30. SNPs private in Lm-2 and Lm-55 were the only ones that consistently occurred in intergenic regions of the genome. The SNP polymorphic information content (PIC) scores ranged from 0.3-0.5, indicative of relatively high polymorphism (Table 3.3). Private allele SNPs were chosen ensuring that their PIC fell within the aforementioned acceptable range. General statistics can be found in Table 3.3 and association analysis of SNPs to the state the isolates were collected from can be found in Figures 3.2-3.5.

Table 3.3 Possible private alleles identified and their location in the *L. maculans* genome

SNP name	SuperContig	Location (bp)	Polymorphic in	Location in genome
SNP 55	SuperContig_2	1120505	Lm-1	Intergenic
SNP 81	SuperContig_13	1254623	Lm-1	Intergenic
SNP 85	SuperContig_13	1424634	Lm-1	Intergenic
SNP 115	SuperContig_8	774901	Lm-2 and Lm-55	Intergenic
SNP 125	SuperContig_10	992433	Lm-2 and Lm-55	Intergenic
SNP 131	SuperContig_6	638539	Lm-2 and Lm-30	Intergenic
SNP 135	SuperContig_11	578821	Lm-2	Intergenic
SNP 137	SuperContig_11	970997	Lm-2 and Lm-55	Intergenic
SNP 140	SuperContig_11	1524610	Lm-1	End of SC
SNP 141	SuperContig_3	95443	Lm-2	In gene similar to peroxisomal membrane protein CBX93615.1
SNP 151	SuperContig_4	346457	Lm-2 and Lm-55	Intergenic
SNP 153	SuperContig_4	873687	Lm-1	Downstream of gene product similar to lipolytic protein G-D-S-L family CBX93216.1
SNP 160	SuperContig_4	1358095	Lm-2	Intergenic
SNP 188	SuperContig_14	264857	Lm-2	Hypothetical protein CDS CBX98019.1
SNP 194	SuperContig_14	1268037	Lm-1	Hypothetical protein CDS CBX98389.1
SNP 198	SuperContig_16	447349	Lm-1	Exon of gene whose product is similar to epoxide hydrolase CBX97267.1
SNP 203	SuperContig_17	350439	Lm-2	Hypothetical protein CDS CBX96839.1
SNP 204	SuperContig_17	523022	Lm-2 and Lm-24	Hypothetical protein CDS CBX96888.1
SNP 205	SuperContig_17	757759	Lm-2	Intergenic

Table 3.4 General statistics of 214 SNPs

(PIC-Polymorphism Information Content; Totals excluding NAs)

SNP name	Total A	Total B	Total	NA	% A	% B	PIC
SNP1	35	56	91	5	38.46	61.54	0.47
SNP2	27	66	93	3	29.03	70.97	0.41
SNP3	35	57	92	4	38.04	61.96	0.47
SNP4	86	10	96	0	89.58	10.42	0.19
SNP5	23	62	85	11	27.06	72.94	0.39
SNP6	13	76	89	7	14.61	85.39	0.25
SNP7	89	5	94	2	94.68	5.32	0.10
SNP8	5	91	96	0	5.21	94.79	0.10
SNP9	36	54	90	6	40.00	60.00	0.48
SNP10	27	68	95	1	28.42	71.58	0.41
SNP11	82	12	94	2	87.23	12.77	0.22
SNP12	72	18	90	6	80.00	20.00	0.32
SNP13	32	54	86	10	37.21	62.79	0.47
SNP14	36	54	90	6	40.00	60.00	0.48
SNP15	8	64	72	24	11.11	88.89	0.20
SNP16	56	31	87	9	64.37	35.63	0.46
SNP17	40	47	87	9	45.98	54.02	0.50
SNP18	42	49	91	5	46.15	53.85	0.50
SNP19	23	67	90	6	25.56	74.44	0.38
SNP20	48	41	89	7	53.93	46.07	0.50
SNP21	33	55	88	8	37.50	62.50	0.47
SNP22	24	63	87	9	27.59	72.41	0.40
SNP23	37	55	92	4	40.22	59.78	0.48
SNP24	83	12	95	1	87.37	12.63	0.22
SNP25	34	58	92	4	36.96	63.04	0.47
SNP26	19	68	87	9	21.84	78.16	0.34
SNP27	30	56	86	10	34.88	65.12	0.45
SNP28	16	75	91	5	17.58	82.42	0.29
SNP29	5	82	87	9	5.75	94.25	0.11
SNP30	35	58	93	3	37.63	62.37	0.47
SNP31	90	5	95	1	94.74	5.26	0.10
SNP32	52	40	92	4	56.52	43.48	0.49
SNP33	37	48	85	11	43.53	56.47	0.49
SNP34	73	23	96	0	76.04	23.96	0.36
SNP35	23	64	87	9	26.44	73.56	0.39
SNP36	80	12	92	4	86.96	13.04	0.23
SNP37	57	33	90	6	63.33	36.67	0.46
SNP38	35	50	85	11	41.18	58.82	0.48
SNP39	32	60	92	4	34.78	65.22	0.45
SNP40	28	67	95	1	29.47	70.53	0.42
SNP41	31	60	91	5	34.07	65.93	0.45

SNP42	29	56	85	11	34.12	65.88	0.45
SNP43	65	28	93	3	69.89	30.11	0.42
SNP44	35	58	93	3	37.63	62.37	0.47
SNP45	45	44	89	7	50.56	49.44	0.50
SNP46	38	52	90	6	42.22	57.78	0.49
SNP47	7	83	90	6	7.78	92.22	0.14
SNP48	9	84	93	3	9.68	90.32	0.17
SNP49	72	20	92	4	78.26	21.74	0.34
SNP50	74	20	94	2	78.72	21.28	0.33
SNP51	20	71	91	5	21.98	78.02	0.34
SNP52	44	47	91	5	48.35	51.65	0.50
SNP53	30	65	95	1	31.58	68.42	0.43
SNP54	77	16	93	3	82.80	17.20	0.28
SNP55	93	3	96	0	96.88	3.13	0.06
SNP56	20	71	91	5	21.98	78.02	0.34
SNP57	23	71	94	2	24.47	75.53	0.37
SNP58	30	59	89	7	33.71	66.29	0.45
SNP59	68	25	93	3	73.12	26.88	0.39
SNP60	74	20	94	2	78.72	21.28	0.33
SNP61	10	75	85	11	11.76	88.24	0.21
SNP62	74	16	90	6	82.22	17.78	0.29
SNP63	63	28	91	5	69.23	30.77	0.43
SNP64	6	89	95	1	6.32	93.68	0.12
SNP65	7	84	91	5	7.69	92.31	0.14
SNP66	46	43	89	7	51.69	48.31	0.50
SNP67	90	6	96	0	93.75	6.25	0.12
SNP68	76	17	93	3	81.72	18.28	0.30
SNP69	33	61	94	2	35.11	64.89	0.46
SNP70	75	20	95	1	78.95	21.05	0.33
SNP71	61	32	93	3	65.59	34.41	0.45
SNP72	88	7	95	1	92.63	7.37	0.14
SNP73	10	84	94	2	10.64	89.36	0.19
SNP74	29	60	89	7	32.58	67.42	0.44
SNP75	15	67	82	14	18.29	81.71	0.30
SNP76	49	43	92	4	53.26	46.74	0.50
SNP77	22	66	88	8	25.00	75.00	0.38
SNP78	46	43	89	7	51.69	48.31	0.50
SNP79	30	59	89	7	33.71	66.29	0.45
SNP80	17	78	95	1	17.89	82.11	0.29
SNP81	3	93	96	0	3.13	96.88	0.06
SNP82	55	31	86	10	63.95	36.05	0.46
SNP83	45	38	83	13	54.22	45.78	0.50
SNP84	57	39	96	0	59.38	40.63	0.48
SNP85	3	81	84	12	3.57	96.43	0.07
SNP86	6	43	49	47	12.24	87.76	0.21

SNP87	47	44	91	5	51.65	48.35	0.50
SNP88	81	14	95	1	85.26	14.74	0.25
SNP89	66	28	94	2	70.21	29.79	0.42
SNP90	35	58	93	3	37.63	62.37	0.47
SNP91	27	59	86	10	31.40	68.60	0.43
SNP92	14	64	78	18	17.95	82.05	0.29
SNP93	16	80	96	0	16.67	83.33	0.28
SNP94	84	8	92	4	91.30	8.70	0.16
SNP95	48	40	88	8	54.55	45.45	0.50
SNP96	84	11	95	1	88.42	11.58	0.20
SNP97	88	7	95	1	92.63	7.37	0.14
SNP98	36	58	94	2	38.30	61.70	0.47
SNP99	23	69	92	4	25.00	75.00	0.38
SNP100	18	67	85	11	21.18	78.82	0.33
SNP101	36	49	85	11	42.35	57.65	0.49
SNP102	36	49	85	11	42.35	57.65	0.49
SNP103	10	85	95	1	10.53	89.47	0.19
SNP104	42	50	92	4	45.65	54.35	0.50
SNP105	67	24	91	5	73.63	26.37	0.39
SNP106	39	55	94	2	41.49	58.51	0.49
SNP107	47	41	88	8	53.41	46.59	0.50
SNP108	5	91	96	0	5.21	94.79	0.10
SNP109	25	66	91	5	27.47	72.53	0.40
SNP110	24	65	89	7	26.97	73.03	0.39
SNP111	42	49	91	5	46.15	53.85	0.50
SNP112	45	48	93	3	48.39	51.61	0.50
SNP113	64	29	93	3	68.82	31.18	0.43
SNP114	22	69	91	5	24.18	75.82	0.37
SNP115	90	4	94	2	95.74	4.26	0.08
SNP116	46	42	88	8	52.27	47.73	0.50
SNP117	5	89	94	2	5.32	94.68	0.10
SNP118	54	39	93	3	58.06	41.94	0.49
SNP119	90	5	95	1	94.74	5.26	0.10
SNP120	47	32	79	17	59.49	40.51	0.48
SNP121	53	38	91	5	58.24	41.76	0.49
SNP122	28	63	91	5	30.77	69.23	0.43
SNP123	37	57	94	2	39.36	60.64	0.48
SNP124	11	84	95	1	11.58	88.42	0.20
SNP125	4	92	96	0	4.17	95.83	0.08
SNP126	16	74	90	6	17.78	82.22	0.29
SNP127	14	76	90	6	15.56	84.44	0.26
SNP128	7	53	60	36	11.67	88.33	0.21
SNP129	80	11	91	5	87.91	12.09	0.21
SNP130	92	4	96	0	95.83	4.17	0.08
SNP131	4	79	83	13	4.82	95.18	0.09

SNP132	49	39	88	8	55.68	44.32	0.49
SNP133	23	70	93	3	24.73	75.27	0.37
SNP134	66	24	90	6	73.33	26.67	0.39
SNP135	3	83	86	10	3.49	96.51	0.07
SNP136	76	16	92	4	82.61	17.39	0.29
SNP137	4	92	96	0	4.17	95.83	0.08
SNP138	91	5	96	0	94.79	5.21	0.10
SNP139	85	11	96	0	88.54	11.46	0.20
SNP140	92	3	95	1	96.84	3.16	0.06
SNP141	3	89	92	4	3.26	96.74	0.06
SNP142	24	70	94	2	25.53	74.47	0.38
SNP143	51	36	87	9	58.62	41.38	0.49
SNP144	40	47	87	9	45.98	54.02	0.50
SNP145	74	21	95	1	77.89	22.11	0.34
SNP146	73	15	88	8	82.95	17.05	0.28
SNP147	92	4	96	0	95.83	4.17	0.08
SNP148	9	86	95	1	9.47	90.53	0.17
SNP149	38	50	88	8	43.18	56.82	0.49
SNP150	65	29	94	2	69.15	30.85	0.43
SNP151	4	88	92	4	4.35	95.65	0.08
SNP152	31	61	92	4	33.70	66.30	0.45
SNP153	3	89	92	4	3.26	96.74	0.06
SNP154	29	58	87	9	33.33	66.67	0.44
SNP155	45	42	87	9	51.72	48.28	0.50
SNP156	19	71	90	6	21.11	78.89	0.33
SNP157	86	7	93	3	92.47	7.53	0.14
SNP158	66	24	90	6	73.33	26.67	0.39
SNP159	70	23	93	3	75.27	24.73	0.37
SNP160	93	3	96	0	96.88	3.13	0.06
SNP161	3	60	63	33	4.76	95.24	0.09
SNP162	74	20	94	2	78.72	21.28	0.33
SNP163	91	5	96	0	94.79	5.21	0.10
SNP164	75	17	92	4	81.52	18.48	0.30
SNP165	69	25	94	2	73.40	26.60	0.39
SNP166	56	35	91	5	61.54	38.46	0.47
SNP167	61	27	88	8	69.32	30.68	0.43
SNP168	69	24	93	3	74.19	25.81	0.38
SNP169	29	63	92	4	31.52	68.48	0.43
SNP170	30	55	85	11	35.29	64.71	0.46
SNP171	39	50	89	7	43.82	56.18	0.49
SNP172	75	17	92	4	81.52	18.48	0.30
SNP173	17	76	93	3	18.28	81.72	0.30
SNP174	41	47	88	8	46.59	53.41	0.50
SNP175	4	88	92	4	4.35	95.65	0.08
SNP176	26	65	91	5	28.57	71.43	0.41

SNP177	17	76	93	3	18.28	81.72	0.30
SNP178	9	79	88	8	10.23	89.77	0.18
SNP179	59	34	93	3	63.44	36.56	0.46
SNP180	33	59	92	4	35.87	64.13	0.46
SNP181	3	69	72	24	4.17	95.83	0.08
SNP182	25	69	94	2	26.60	73.40	0.39
SNP183	71	17	88	8	80.68	19.32	0.31
SNP184	22	69	91	5	24.18	75.82	0.37
SNP185	48	42	90	6	53.33	46.67	0.50
SNP186	81	14	95	1	85.26	14.74	0.25
SNP187	22	63	85	11	25.88	74.12	0.38
SNP188	3	92	95	1	3.16	96.84	0.06
SNP189	54	35	89	7	60.67	39.33	0.48
SNP190	13	79	92	4	14.13	85.87	0.24
SNP191	15	76	91	5	16.48	83.52	0.28
SNP192	70	21	91	5	76.92	23.08	0.36
SNP193	7	85	92	4	7.61	92.39	0.14
SNP194	3	93	96	0	3.13	96.88	0.06
SNP195	39	49	88	8	44.32	55.68	0.49
SNP196	66	25	91	5	72.53	27.47	0.40
SNP197	7	80	87	9	8.05	91.95	0.15
SNP198	93	3	96	0	96.88	3.13	0.06
SNP199	41	50	91	5	45.05	54.95	0.50
SNP200	13	83	96	0	13.54	86.46	0.23
SNP201	53	39	92	4	57.61	42.39	0.49
SNP202	40	47	87	9	45.98	54.02	0.50
SNP203	3	93	96	0	3.13	96.88	0.06
SNP204	4	82	86	10	4.65	95.35	0.09
SNP205	92	3	95	1	96.84	3.16	0.06
SNP206	73	20	93	3	78.49	21.51	0.34
SNP207	75	18	93	3	80.65	19.35	0.31
SNP208	32	57	89	7	35.96	64.04	0.46
SNP209	60	33	93	3	64.52	35.48	0.46
SNP210	76	19	95	1	80.00	20.00	0.32
SNP211	45	46	91	5	49.45	50.55	0.50
SNP212	33	57	90	6	36.67	63.33	0.46
SNP213	19	72	91	5	20.88	79.12	0.33
SNP214	25	60	85	11	29.41	70.59	0.42

3.3.3 Population Analysis

No significant correlation was observed in the data between the SNPs and the isolate collection site, the stubble cultivar, resistance complement of the cultivar or the year the isolates were collected (Table 3.1). A low level of association was observed between SNP73 on SC 12 and *B. napus* (as the stubble species) compared to *B. juncea* cultivars as the stubble species (Figure 3.1). No association to state (Figures 3.2-3.5), stubble cultivar, year collected or place collected could be found (Data not shown).

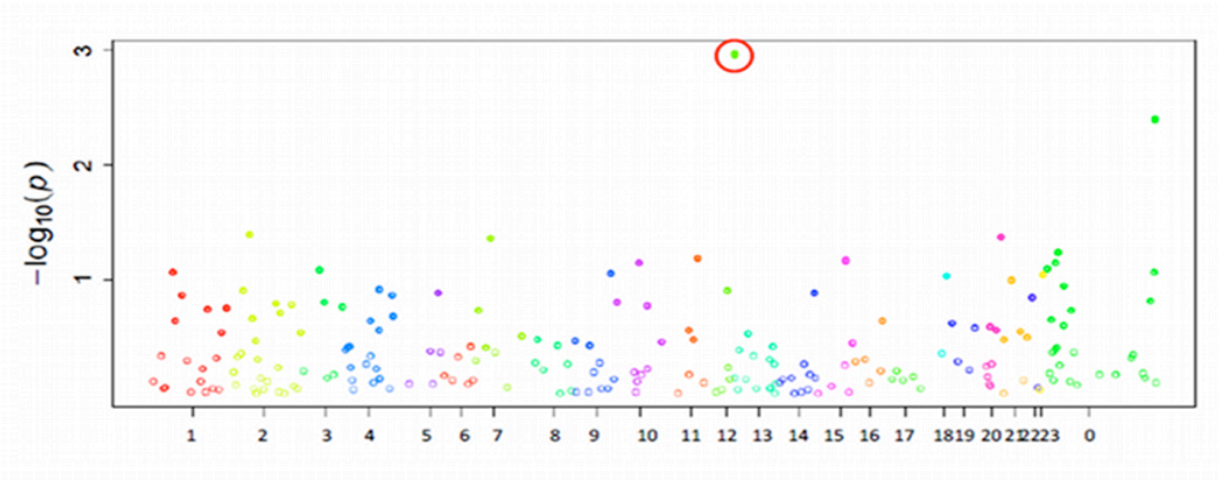

Figure 3.1 Manhattan plot of all isolates showing association between SNPs and *B. napus* species (SNP 73 circled in red; x-axis supercontigs; y-axis –log10 values of association between stubble species and SNPs)

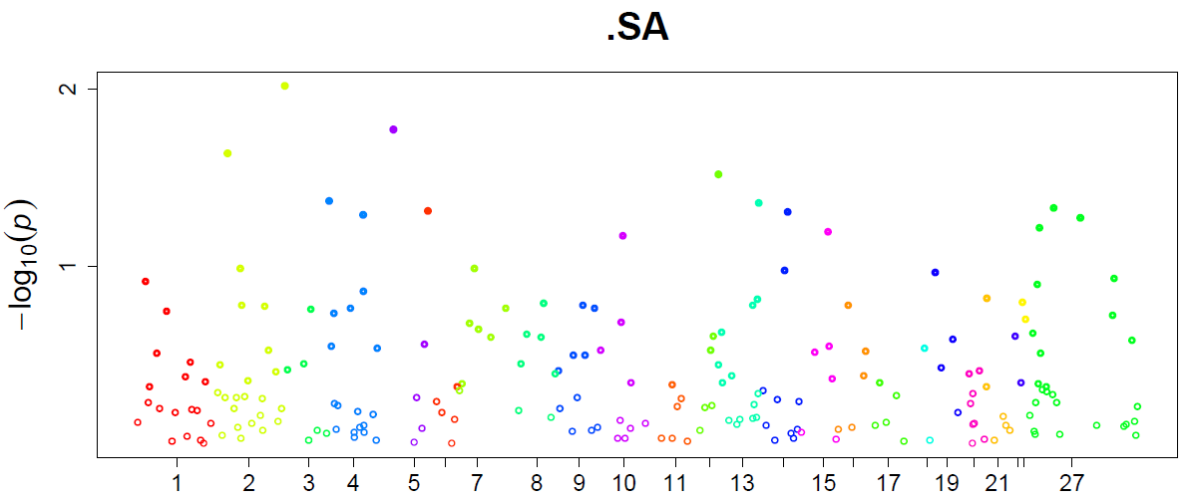

Figure 3.2 Manhattan plot of all isolates showing association between SNPs and South Australia (X-axis: Supercontigs (27= SC0); y-xis $-\log_{10}$ values of association between state of collection and SNPs)

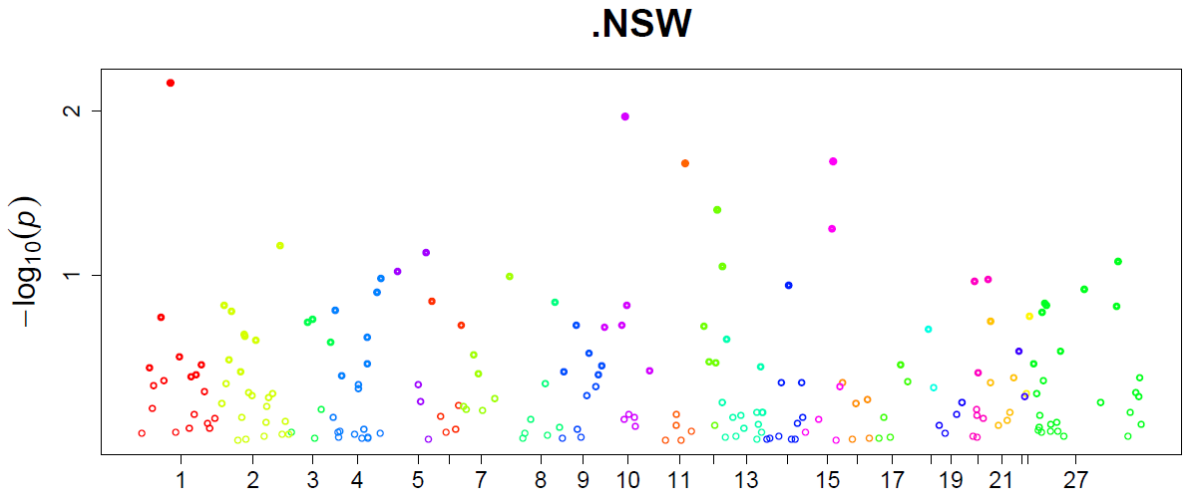

Figure 3.3 Manhattan plot of all isolates showing association between SNPs and New South Wales (X-axis supercontigs (27= SC0); y-xis $-\log_{10}$ values of association between state of collection and SNPs)

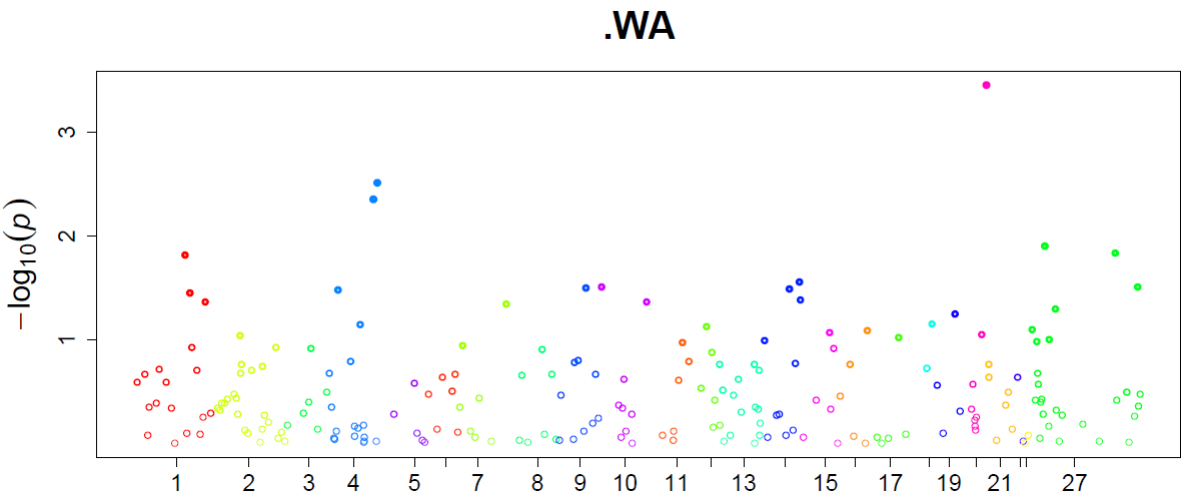

Figure 3.4 Manhattan plot of all isolates showing association between SNPs and Western Australia (X-axis supercontigs (27= SC0); y-xis $-\log_{10}$ values of association between state of collection and SNPs)

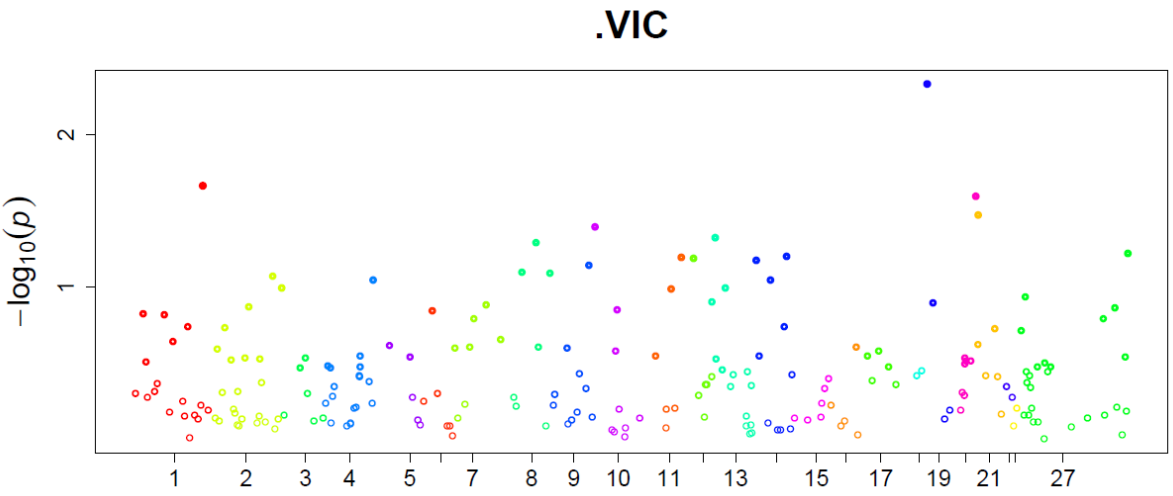

Figure 3.5 Manhattan plot of all isolates showing association between SNPs and Victoria (X-axis supercontigs (27= SC0); y-xis $-\log_{10}$ values of association between state of collection and SNPs)

3.3.4 Phylogenetic Tree

In order to visually analyse relationships between the isolates, a dendrogram was generated. The isolates used for SNP prediction, Lm-1 and Lm-2 are on separate clades as can be discerned from the resulting phylogenetic tree (Figure 3.6). Based on this analysis, a number of isolates appeared to be genetically identical. DNA replicates used in the assay were noted to be the same (Table 3.2).

From the dendrogram it can be seen that the isolate Lm-55 was most similar to Lm-2 (Figure 3.6). Both isolates were collected in the same year (2006) and the same region of the state (Victoria) but from different stubble cultivars of *B. napus*. Isolates Lm-7 and Lm-28 were also seen to be similar. However, Lm-7 was collected in 2004 and Lm-28 in 2003 from different stubble cultivars (Grace and ATR-Beacon respectively) and from different states (Wonwondah in Victoria and Moyhall in South Australia respectively, both close to a common border about 120 km apart). Boxes 1-8 denote isolates that group locally based on year collected, site of collection, state, stubble species and/or stubble cultivar (Figure 3.6), the details for which are seen in Table 3.1. The majority of clusters were from Victoria and South Australia. The most common groupings were based on state and stubble cultivar.

Some differences between certain isolates grouping together were also noticeable. Lm-16 and Lm-39 were isolated in different years (2004, 2005), from different stubble cultivars (ATR-Beacon, Surpass501TT) and in different places (Bordertown, SA and Mt. Barker, WA). The same differences could be seen for Lm-41 and Lm-52, Lm-32 and Lm-45, Lm15 and Lm-20, Lm-4, Lm-8 and Lm-40, Lm-6 and Lm-18 and Lm-21 and Lm-27. Therefore, small groupings based on the parameters listed in Table 3.1 were seen throughout the tree along with larger differences. No other patterns could be elicited from the positioning of sub-clades in this tree.

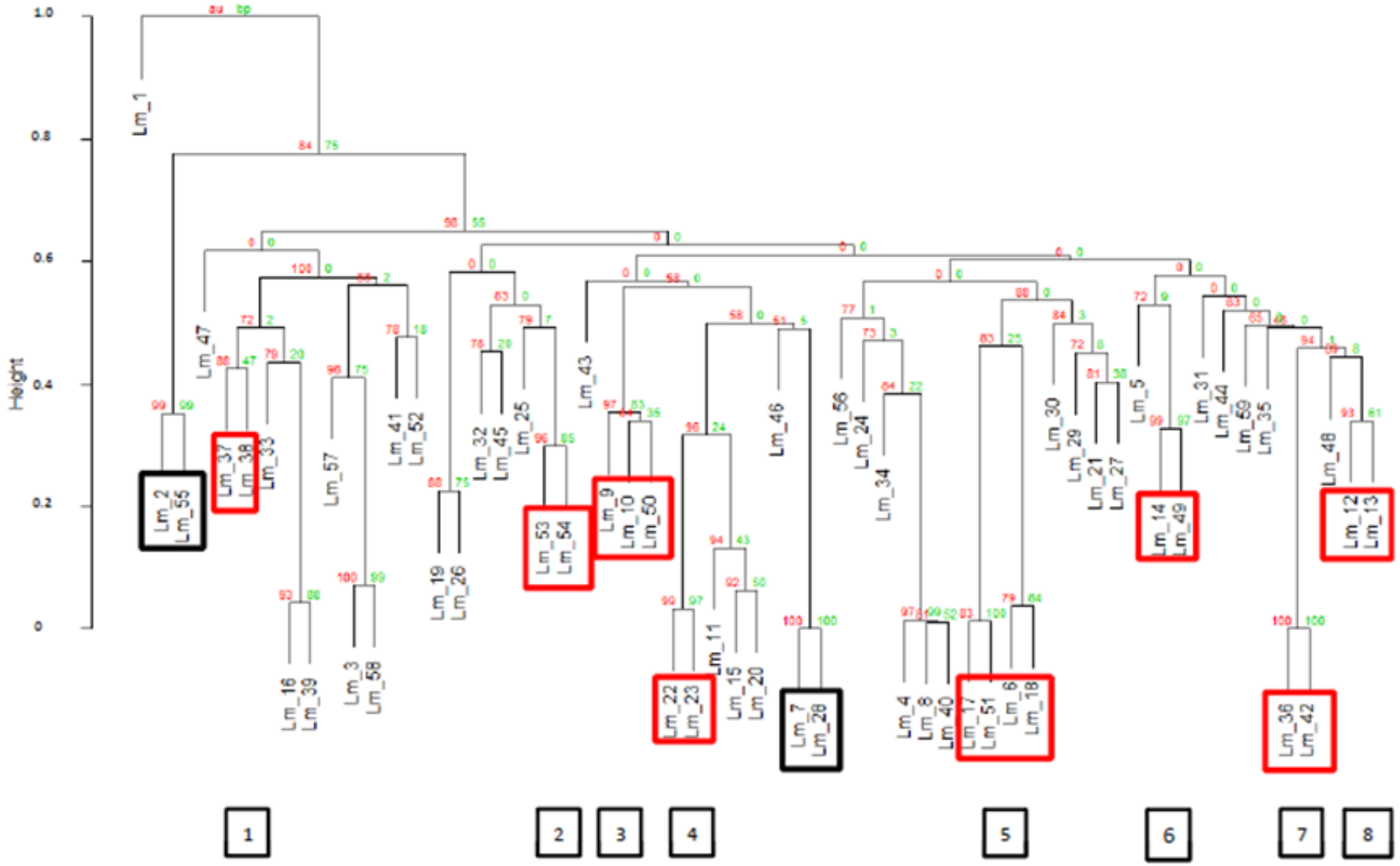

Figure 3.6 Phylogenetic tree based on all isolates

(Bootstrap values: red-Approximately Unbiased (AU) p-values calculated by multiscale bootstrap resampling; green-Bootstrap Probability (BP) p-values calculated by normal bootstrap resampling; AU values >95 % strongly supported by data (Suzuki and Shimodaira, 2006); Replicates not shown; Left to right red boxes 1-8; Black boxes indicate similar isolates)

3.3.5 Principal Component Analysis (PCA)

The PCA primarily showed a random distribution of isolates along the two principal component axes (Figure 3.7). Isolates Lm-1 and Lm-2 were completely different to each other, as expected based on their use for identification of polymorphic SNPs for designing the assay. Genetically identical isolates were plotted at the same point or grouped together in the same area, validating the results of the dendrogram (Figure 3.6). Isolates Lm-36 and Lm-42 clustered together. Both these isolates were collected in 1988 in Victoria, from different locations (Table 3.1). No other conclusive correlations could be elucidated.

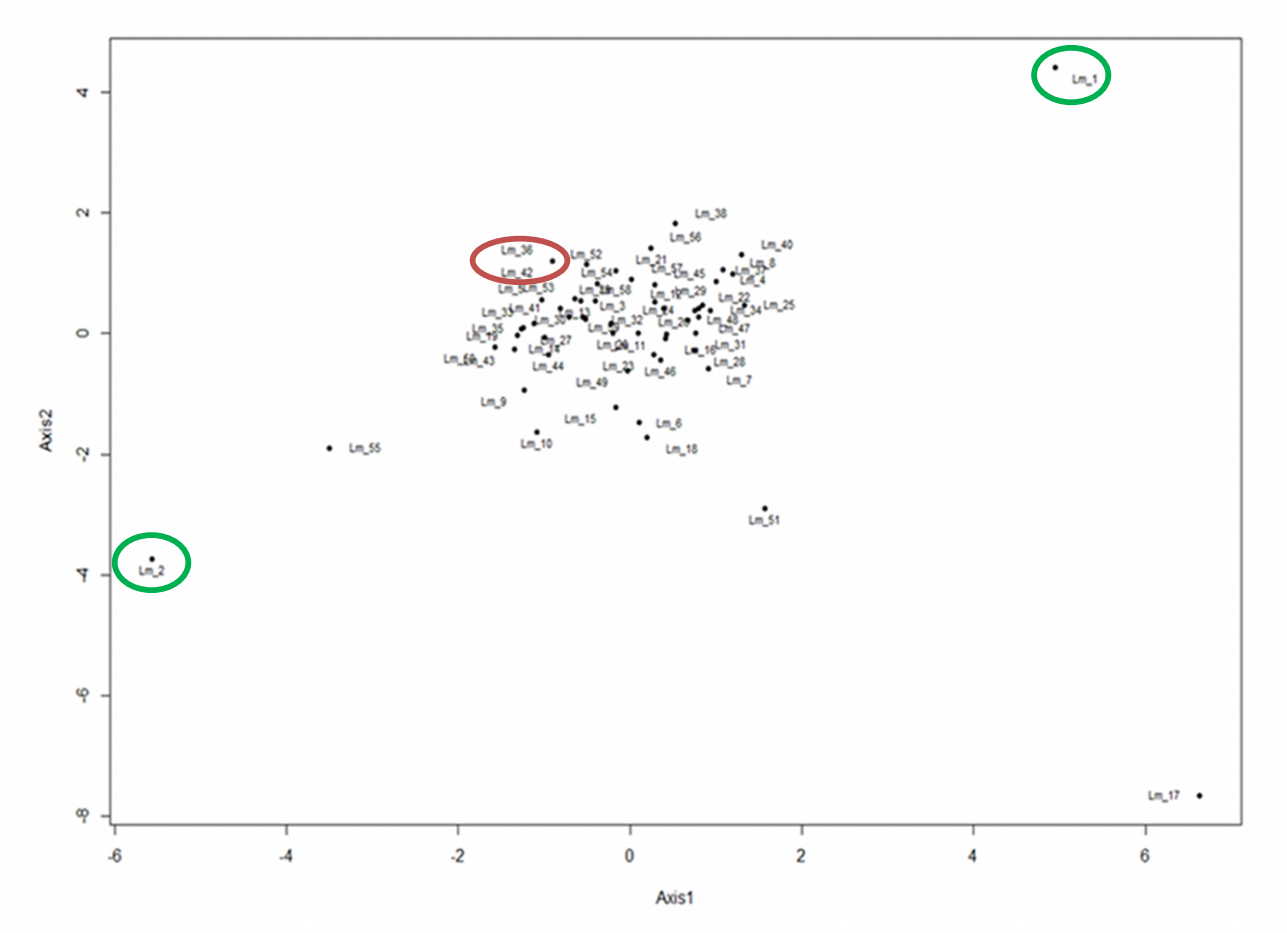

Figure 3.7 PCA displaying correlation between Australian *L. maculans* population isolates across 214 SNP loci isolates

(Axis1-Principal Component 1, Axis2-Principal Component 2; SNP prediction based on Lm-1 and Lm-2-highlighted by green circles; Lm-36 and Lm-42 highlighted by orange circle; replicates not shown)

3.3.6 Linkage disequilibrium (LD)

The sub-data set of 193 SNPs, after elimination of 21 SNPs considered private alleles, was analysed to measure pairwise LD between SNPs. The heatmap in Figure 3.8 displays results of the LD calculation. Based on the R^2 values displayed in the heatmap, little significant LD was observed in the *L. maculans* population. R^2 =1 indicates no recombination, and thus high LD, and R^2 =0 indicates considerable recombination thus no LD. Only 0.83% of p-values associated with pairwise comparisons were significant. Furthermore, the heatmap failed to show any noticeable patterns or blocks of LD.

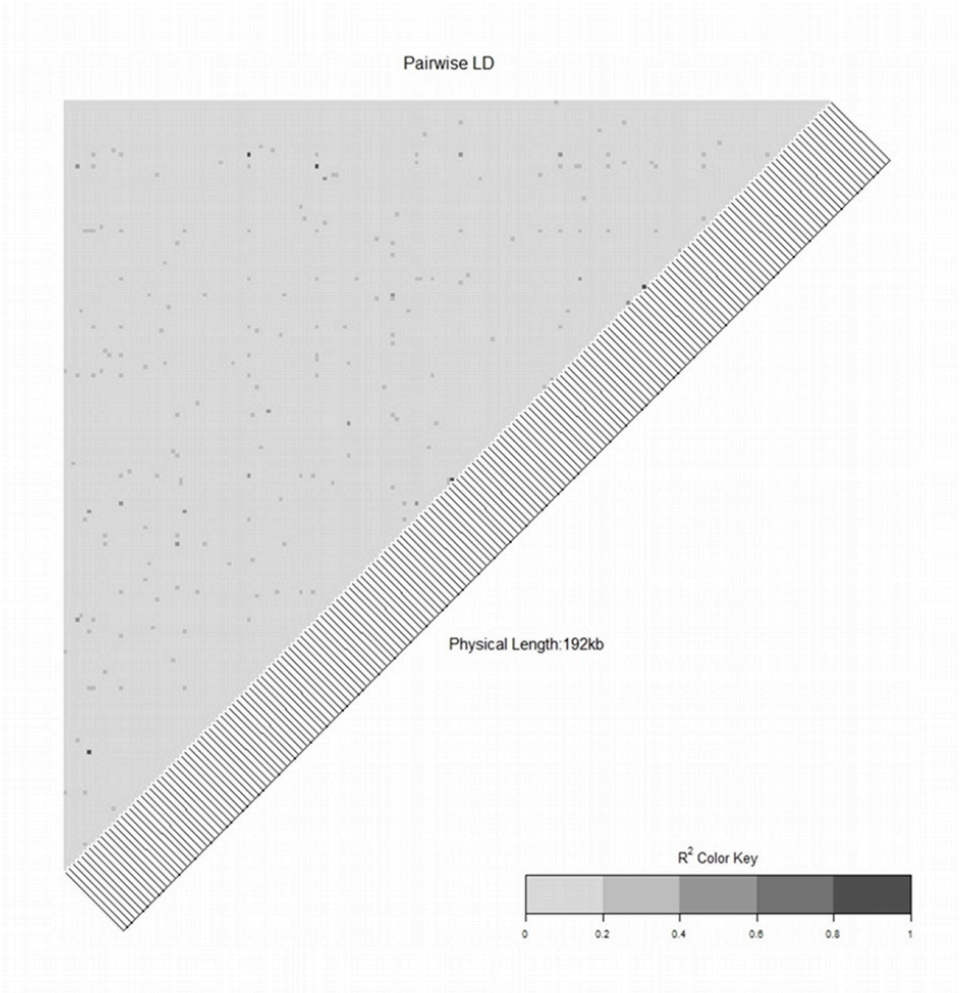

Figure 3.8 Heatmap displaying Linkage disequilibrium in Australian *L. maculans* population isolates across 193 SNP loci

(R^2=0 considerable recombination; R^2=1 no recombination)

3.4 Discussion

The isolates used in this study were collected from all around Australia. The absence of any obvious genetic variance specific to a certain parameter suggests that the Australian *L. maculans* isolates comprise a single population, but that some possible subpopulations localised to the state or site of collection do occur. Cultivar stubble may assist in maintaining the large population size (Travadon et al., 2011; Daverdin et al., 2012). No noticeable patterns or haplotypes were detected in the data, suggesting that this fungus evolves rapidly under selection pressure from the host. Sexual reproduction in this pathogen facilitates the production of ascospores which is its primary inoculum (Rouxel and Balesdent, 2005). Human transport may aid in transporting infected material to different regions (Travadon et al., 2011). This in turn leads to random mating between isolates from different regions, creating genetic variance at the avirulence loci and assisting the pathogen to overcome host resistance (Dilmaghani et al., 2012).

It is expected that the population of a sexual reproductively active pathogen will be panmictic. In a panmictic population, members may interact with one another at random, which creates extensive recombination and genetic diversity (Polk and Peek, 2010). Previous *L. maculans* population studies have reached the same conclusion of panmixia (Barrins et al., 2004; Travadon et al., 2011). We assumed the null hypothesis of a panmictic blackleg population across Australia while conducting population analyses, which was supported by the results of the manhattan plots, dendrogram, PCA and LD analysis.

Overall, the phylogenetic tree and PCA analysis suggested that *L. maculans* possesses a high evolutionary potential, indicative of populations able to overcome genetic resistance (McDonald and Linde, 2002). Large population size, high rate of mutation, high genotype flow and mixed reproduction all confer high evolutionary potential to the pathogen, putatively enabling it to overcome host genetic resistance (McDonald and Linde, 2002). The length of the tree branches indicates genetic similarity between isolates. Based on this, isolates Lm-28 and Lm-7 appear to be genetically identical with 78% shared alleles (22% missing values). These isolates have been collected from Moyhall, South Australia in 2003 and Wonwondah, Victoria in 2004 respectively. Van de Wouw et al. (2010) also used isolates Lm-6, Lm-17, Lm-18 and Lm-51 for genotyping at the *AvrLm1* and *AvrLm6* loci. They classified Lm-18 and Lm-51 as haplotype 24, Lm-17 as haplotype 10 and Lm-6 as haplotype 4 based on their *Avr* genotype. These haplotypes displayed a completely different association in their study as compared to the dendrogram. This highlights the

efficacy of using a large number of SNPs chosen from across the genome to classify the relationships between isolates.

3.4.1 Phylogenetic tree and PCA

Some local groupings based on state, year collected, stubble species, stubble cultivar and/or site of collection were also observed. Isolates in Box 1-8 (Figure 3.6) all clustered for certain parameters, with each cluster group sharing a single state of collection. Local groupings such as these may indicate the presence of small clonal subpopulations of this pathogen within Australia, such as were observed by Dilmaghani et al. (2013) on *B. oleracea* in Mexico. Clusters in our study were spread across the tree, indicating genetic differences between these possible subpopulations. Different conditions particular to each state such as weather, cultivars grown and stubble resistance all cumulatively affect the evolution of this pathogen. Therefore, it may be that conditions particular to each state promote asexual reproduction rather than sexual reproduction leading to less diversity within each subpopulation. Dilmaghani et al. (2013) attributed clonality such as this to moving the pathogen from its native biogeographic range, loss of a mating-type by mutation and culture conditions conducive to large-scale dispersal of conidia. This conclusion was made based on the presence of high linkage disequilibrium. Certain isolates like Lm-16 and Lm-39 also clustered together but were vastly different in the parameters associated with them. It is known that human movement transports and introduces infected seed and plant material from one area to another (Dilmaghani et al., 2012) thereby mixing and changing the population, further attributing to its panmictic nature. The bootstrapping of the phylogenetic tree also supports the theory of a randomly interacting mixed population. Bootstrapping values for the main branches were <95%, which indicates low confidence in the hierarchical cluster analysis. On the other hand, most bootstrapping values within each box were >95%, indicating that they were strongly supported by the data (Suzuki and Shimodaira, 2006). An overall analysis of the tree yielded no particular association to any other parameter.

The PCA results also displayed a random positioning of isolates. Certain isolates clustered together, such as Lm-36 and Lm-42 isolates in Box 7, validating the dendrogram and also supporting panmixia. These findings could be attributed to the high evolutionary potential of *L. maculans*. Spore dispersal also plays a role in increasing gene flow and generating a random mix of isolates across the population (Travadon et al., 2011). More samples from each region and year will need to

be collected and examined to investigate the possibility of clonal sub-populations of *L. maculans* within Australia.

3.4.2 Linkage Disequilibrium

We hypothesised that a possible cause of LD in this population could be selection of *AvrLm* genes, due to their impact on host plant infection. This would lead to loci in the selected region segregating with each other more often than expected by chance and can be visualised as blocks on the LD heatmap. However, we failed to notice any such patterns. As the rate of recombination between loci increases, there is a greater chance of linkage equilibrium in the population, decreasing LD. Populations that are constantly recombining and have a high cross-over rate will show little LD (McVean, 2008). Xu (2006) stated that only 5% of locus-pairs have significant observed association to those expected in a completely panmictic population. Our LD analysis showed 0.83% of p-values associated with pairwise comparisons, to be significant. SNP73, which was seen to be significantly associated with *B. napus* cultivars and was located near *Avr4-7*, did not display significant LD. The SNP and the gene on SC12 of the *L. maculans* genome are 253.6 kb apart and the GC content of the traversing region is 45.2%. Parlange et al. (2009) reported two PCR markers on the border of the *AvrLm4-7* locus; the GC content between those markers was 35.2%. It has been concluded by Rouxel et al. (2011) that recombination in the *L. maculans* genome occurs more frequently within GC-rich regions than between GC-rich regions. Therefore, we believe that recombination events in the GC-rich region between the SNP and the gene may have impacted the association between them in *B. napus* cultivars. The number of samples isolated from *B. napus* cultivars may also be too small to clearly display this association on the heatmap in the form of LD. Furthermore, the majority of Australian cultivars that have been genotyped contain *Rlm4* and therefore there has been strong selection pressure at the *AvrLm4-7* locus for a number of years in Australia (Marcroft et al., 2012). However, detailed studies comprising more isolates derived from *B. juncea* isolates will need to be conducted to confirm the validity of this association to *B. napus* cultivars. In the future, examining associations between SNPs and *Avr* genes in AT-rich regions such as these may prove fruitful in analysing the evolution of avirulence genes.

The results from this SNP genotyping assay successfully validated the work conducted by Zander et al. (2013) using the same SNP resource. The SNP prediction supported transferability of SNPs for use in the GoldenGate assay for the chosen SNPs. Stringent clustering criteria (ensuring that all SNPs visually separated into either the ‘A’ group or the ‘B’ group) yielded 214 SNPs, which were used for subsequent data analysis. Poor clustering could be a result of additional SNPs in the

flanking regions of the predicted SNPs which can be resolved in the future by using more isolates for SNP discovery. Private alleles relating to five particular isolates (Lm-1, Lm-2, Lm-24, Lm-30 and Lm-55) were found. Lm-2 and Lm-55 were isolated from the same year and site of collection in Victoria. Alleles present only in the isolates used for SNP prediction (Lm-1 and Lm-2) are due to ascertainment bias of using these for the SNP prediction. Ascertainment bias is introduced because of the method used for SNP discovery (Albrechtsen et al., 2010), which in this case, used two isolates (Lm-1 and Lm-2) to predict SNPs. Ascertainment bias can be corrected in the future by using more isolates for SNP prediction. Being closely related and hence sharing sequence similarity might explain the four alleles that were found only in Lm-2 and Lm-55. The two main clades on the phylogenetic tree separated the isolates Lm-1 and Lm-2. Replicates used in this assay were seen to be genetically identical except for missing values. This was validated by the phylogenetic tree and PCA output, which confirmed the reproducibility of this assay.

The purpose behind conducting a large-scale genotyping assay was to understand the genetic diversity in the Australian *L. maculans* population. Our cumulative analysis of these results supports our null hypothesis of a panmictic Australian *L. maculans* population with possible regional clonality. This contrasts with the conclusions of Hayden et al. (2007) who found two genetically distinct eastern and western blackleg populations in Australia using six microsatellite and two minisatellite markers in 513 isolates collected over two years. The study also found 85% difference within the 13 subpopulations identified and 10% difference between the coasts. However, our results concur with the panmictic conclusion of Travadon et al. (2011). This study analysed 29 field populations of French *L. maculans* isolates using minisatellite markers and also found low genetic differentiation within populations. Further study concentrated on sampling from each region will help provide insights into this theory.

3.5 Conclusion

Overall, the high rate of sexual reproduction, ability of the pathogen to survive on stubble for long periods of time and random mating between isolates likely assist in maintaining a large blackleg population in Australia. This combined with its high evolutionary potential enables it to overcome host resistance quickly and cause infection leading to wide-spread crop losses. It is therefore imperative to attempt to restrict the population size of this pathogen using existing methods such as stubble management and introgression of resistance genes in *Brassica* species. Future work involves identifying new disease-associated genes, which will help in developing novel strategies to further control this devastating pathogen.

4 Gene loss analysis in isolates of *Leptosphaeria maculans*

4.1 Introduction

Recent whole genome sequencing of filamentous plant pathogens has revealed a trend towards larger genome sizes owing to repetitive DNA content (Raffaele and Kamoun, 2012). These genomes also contain highly adaptable, plastic regions, usually repeat-rich, comprising of non-coding DNA. However, genome size and architecture are often also affected by occurrences such as gene loss, intron loss and low transposon composition (Raffaele and Kamoun, 2012).

Loss of genes is one of the many methods used by an organism to evolve and adapt. As per the "knockout rate" theory, genes whose absence has a greater effect on the fitness of the organism tend to be conserved (Krylov et al., 2003). Therefore, genes which are essential to the fitness of the organism do not evolve as quickly as non-essential genes. It is long known that the genomes of filamentous pathogens are shaped by their interaction with the host (Raffaele and Kamoun, 2012). This may be further influenced by the presence of repeat elements, host-driven selection pressure and illegitimate recombination events. Furthermore, lifestyle, mating and environmental factors also play a role in maintaining or dispensing of necessary versus unnecessary genes in the genome.

Here we focus on the genome of the phytopathogen *Leptosphaeria maculans* that causes disease on canola (*Brassica napus*). The aim of this chapter was to use the gene loss method (Golicz et al., 2015) to analyse the effect loss of different genes can have on genome structure, lifestyle and infection process of the pathogen. Based on gene annotations in the reference genome (Rouxel et al., 2011) and resequencing data, the gene loss tool used here calculates the fraction coverage of exons of a gene and places them into five different categories, one being [not covered] or lost (Golicz et al., 2015). This method can be used to identify disposable genes in the blackleg genome, as well as regions under strong selection pressure.

4.1.1 Features of the blackleg genome

As noted by Grandaubert et al. (2014b) in comparison to its close relatives (*L. maculans 'lepdii'* genome size 31.5 Mb*, L. biglobosa 'brassicae'* genome size 31.8 Mb*, L. biglobosa 'canadensis'* genome size 30.2 Mb and *L. biglobosa 'thlaspii'* genome size 32.1 Mb), *L. maculans 'brassicae'* has the largest genome size of ~45 Mb which was attributed to its sizeable composition of transposable elements (TE). The genome of *L. maculans* was found to comprise of 32.5% of TEs as

compared to the other genomes analysed, whose composition ranges from 2.7% to 4.0% (Grandaubert et al., 2014b). Rouxel et al. (2011) describe a bipartite structure of the genome that comprises of alternating AT-rich and GC-rich blocks. These isochore-like blocks are contrasting in their TE and gene content. The GC-rich blocks are gene-rich, containing 95% of the predicted genes but have almost no TE content, whilst the AT-rich blocks are riddled with TE mosaics and contain 5% of total predicted genes in the genome. These AT-blocks are preferential niches for hosting effector genes and secondary metabolite clusters and are highly plastic due to invasion by TEs. It has been suggested that hosting effector genes involved in infection in plastic regions of the genome enables the pathogen to readily dispose of them when needed (Grandaubert et al., 2014b).

Other features of the *L. maculans* genome include pre-meiotic repeat-induced point (RIP) mutations. This repeat-inactivated mechanism causes C to T and G to A nucleotide base changes (Rouxel et al., 2011). RIP mutations are thought to play a role in creating diversifying effects in genes present in repeat-rich regions, such as AT-blocks. It is also speculated that since effector genes are hosted in repeat-rich regions, the pathogen employs RIP mutations to evolve and overcome host recognition. A gene may be RIP affected to varying degrees based on the need to diversify sequence but retain functionality, or it may completely degenerate the sequence of the effector gene (Rouxel et al., 2011).

According to the gene-for-gene interaction or effector triggered immunity, certain pathogen avirulence genes are recognised by their corresponding host resistance genes (Ansan-Melayah et al., 1998; Dangl and Jones, 2001). In order to avoid host recognition, the pathogen diversifies its effector genes through varies methods including RIP mutations, non-synonymous point mutations, effects of transposons and deletion of the locus containing the gene (Fudal et al., 2009). The latter is the method used by the pathogen for overcoming resistance to *Rlm1* and *Rlm6*.

Based on this prior knowledge, we hypothesised that the SGSGeneLoss method will help us identify novel avirulence genes that are being lost in isolates compared to the reference. We also aimed to use this tool to identify necessary versus disposable genes in the blackleg genome and characterise regions undergoing selective pressure.

4.2 Methods

4.2.1 Sample growth and tissue harvest

Samples were grown and tissue was harvested as per the methods listed in Sections 2.1 and 2.3 of Chapter 2. The isolates used in this study were D1, D2, D3, D4. D5, D6, D8, D9 and D12 (Refer to Chapter 3, Table 3.1 for details).

4.2.2 DNA extraction

DNA extraction of samples was performed as per the protocol outlined in section 2.4 of Chapter 2. All samples were extracted using the 96 Plant DNeasy kit (Qiagen, 2015) according to the manufacturer's instructions.

4.2.3 Quantification and quality analysis

Sample quality was assessed and quantified as per methods detailed in section 2.4 of Chapter 2

4.2.4 PCR validation

PCR validation of *AvrLm6* in isolate D8 was performed using two sets of primers (Table 4.1).

Table 4.1 PCR primers used for *AvrLm6* locus validation

(F-Forward, R-Reverse)

Primer Name	Primer sequence	Product size (bp)	Reference
1-F	AAACGGCACTATTACGAAAA	116	(Fudal et al., 2007)
1-R	GATTAGGCGAGAAGCAAGT		
2-F	TCAATTTGTCTGTTCAAGTTATGGA	667	(Van de Wouw et al., 2014)
2-R	CCAGTTTTGAACCGTAGAGGTAGCA		

Each PCR reaction contained 1 µl of sample DNA, 1 µl of each primer in each pair, 12.5 µl of EmeraldAmp® MAX HS PCR Master Mix (Takara Clontech) and 9.5 µl of dH_2O, made up to a final volume of 25 µl. Thermocycling conditions were as follows- 94 °C for 4 minutes, then 34 cycles of 94 °C for 30 seconds, 58 °C for 30 seconds and 72 °C for 1 minute. This was followed by extension at 72 °C for 10 minutes and ultimately the sample was kept at 15 °C. 10 µl of each PCR

product was run on a 1% w/v agarose gel (See Section 2.5.1) with 3 µl of ladder to visualise the product.

4.2.5 Library preparation and sequencing

Next generation sequence libraries for the isolates used in this study were made using either the TruSeq (D1, D2) or TruSeq Nano (D3, D4, D5, D6, D8, D9, D12) kits as per the manufacturer's instructions (Illumina Inc). D1 and D2 which were sequenced on the Illumina GAIIx platform and the remaining were sequenced on the Illumina HiSeq. A minimum coverage of 10x was recommended to produce accurate results using this tool (Golicz et al., 2015).

4.2.6 Reference genome

This method requires a reference genome (JN3 or v23.1.3), which was obtained from Rouxel et al. (2011).

4.2.7 Gene loss analysis

As detailed in Golicz et al. (2015) SGSGeneLoss was used to conduct gene loss analysis on *L. maculans*. It calculates 'horizontal' (*frac_exons_covered*) and 'vertical' (*cov_cat*) read coverage for every annotated gene. The '*frac_exons_covered'* ranges from 0-1 and is the "proportion of exon regions of a gene covered by mapped reads" (Golicz et al., 2015). '*cov_cat'* comprises coverage categories 0; >0× ≤ 10×; >10× ≤ 20×; >20× ≤ 40×; >40× ≤ 70×; >70× and is the average vertical coverage depth of a gene.

In this study, genes were labelled as 'lost' if <0.05 fraction of the exons was covered by reads. In order to visualise lost vs conserved genes, genes were divided into the following categories- [70+] [40-70], [20-40], [10-20] and [not covered].

4.2.8 Gene annotation

Blast2Go (Conesa et al., 2005) was used to obtain functional annotations of predicted genes which were first obtained from a non-redundant NCBI database. Then, based on top BLAST hits for each gene, Blast2Go assigned Gene Ontology (GO) terms to them using the database 'Public database b2g_sep13'.

4.2.9 RNA Seq data

RNA seq data was obtained from Lowe et al. (2014). Read data is hosted on the Sequence Read Archive (NCBI) under BioProject accession number SRP035525.

4.3 Results

Overall the results indicated varied gene loss in each isolate analysed, as compared to the reference genome. Gene loss was more concentrated on two particular supercontigs, as compared to the rest of the genome. Our analysis also led us to the discovery of a potential secondary metabolite cluster and novel candidate avirulence genes (detailed in chapter 5).

4.3.1 Sequencing

We obtained sequence read coverage ranging from ~51.5x to ~266.3x per isolate, with an average estimated coverage of ~108 x (Table 4.2). Length of reads for all isolates was 100 bp.

Table 4.2 Whole genome sequencing of isolates

(bp-Base Pair; Insert Size-Distance between two paired reads; Total #bp-Total number of reads obtained after sequencing; Estimated coverage-The number of times (x) the total #bp can cover the entire *L. maculans* genome)

Name	Read Length (bp)	Estimated Insert Size	Number of Paired Reads	Total # bp	Estimated Coverage
D1	100	200	59,493,424	12,017,671,648	266.3x
D2	100	250	11,492,603	2,320,000,000	51.5x
D3	100	580	17,778,999	3,555,799,800	78.8x
D4	100	500	22,401,329	4,480,265,800	99.3x
D5	100	500	22,935,182	4,587,036,400	101.7x
D6	100	500	19,967,228	3,993,445,600	88.5x
D8	100	500	28,463,112	5,692,622,400	126.2x
D9	100	530	16,293,997	3,258,799,400	72.2x
D12	100	520	19,507,705	3,901,541,000	86.5x

4.3.2 Gene loss statistics

A total of 92 genes were found to be lost, as compared to the reference, across all 76 SCs of *L. maculans* in the nine isolates analysed (Figure 4.1), which makes up 0.7% of predicted genes in the genome. This includes fifteen genes which were lost in common amongst all isolates (See Section 4.3.7). Approximately 15% of gene loss (Figure 4.2) occurred on SC_12, which also comprised seven of the fifteen commonly lost genes. This is a loss of 14 genes out of the total 92 reported lost genes. In addition, approximately 24% of genes were lost from SC_13 (Figure 4.2), where twelve lost genes were clustered in a ~27 kbp region (Table 4.4). This made these two SCs the region of maximum gene loss in the genome. Clustered gene loss was only observed on SC_13 and although SC_12 displayed most gene loss, genes were lost singularly or in pairs simultaneously placed along the genome. A total of 28 genes were lost uniquely across all isolates, with the highest number of unique genes lost (six) on SC_13 (See section 4.3.5). Only one gene was lost on SC_27, whilst SC_10, SC_11, SC_24-SC_26, SC_28-SC_76 and SC_30 did not show any gene loss. A total of 33 of 37 genes previously identified as lost in isolates 21 and 41 (Table 4.3; Golicz et al., 2015) were also lost in our dataset.

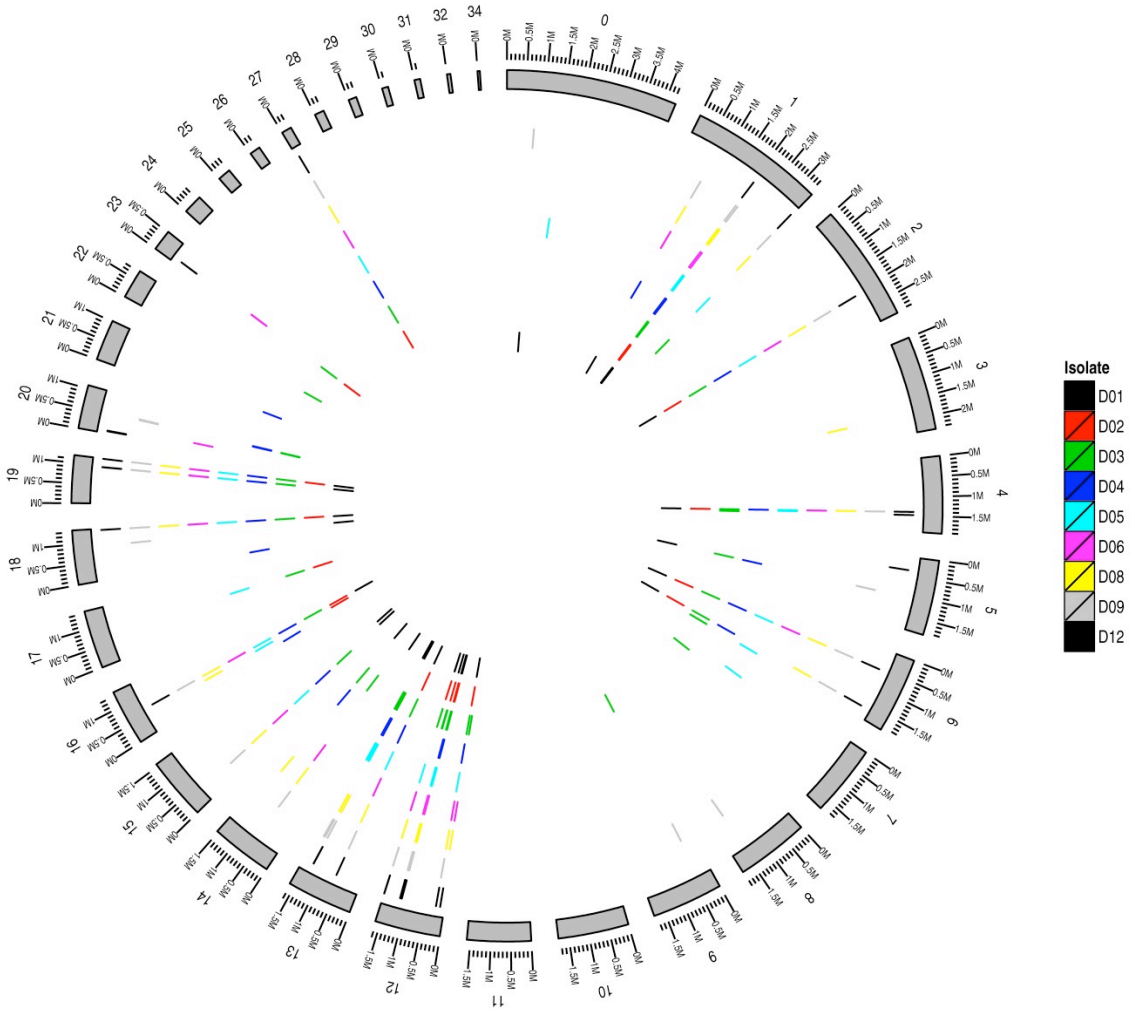

Figure 4.1 Pictorial representation of total gene loss in the *L. maculans* genome

(The outer ring of numbers represents supercontigs of the *L. maculans* genome; the middle ring denotes position information from 0 Mb up to 4 Mb, based on the length of each supercontig; each grey block corresponds to a supercontig containing annotated genes; the innermost rings display position of genes loss per isolate (coloured lines-see key) on a supercontig)

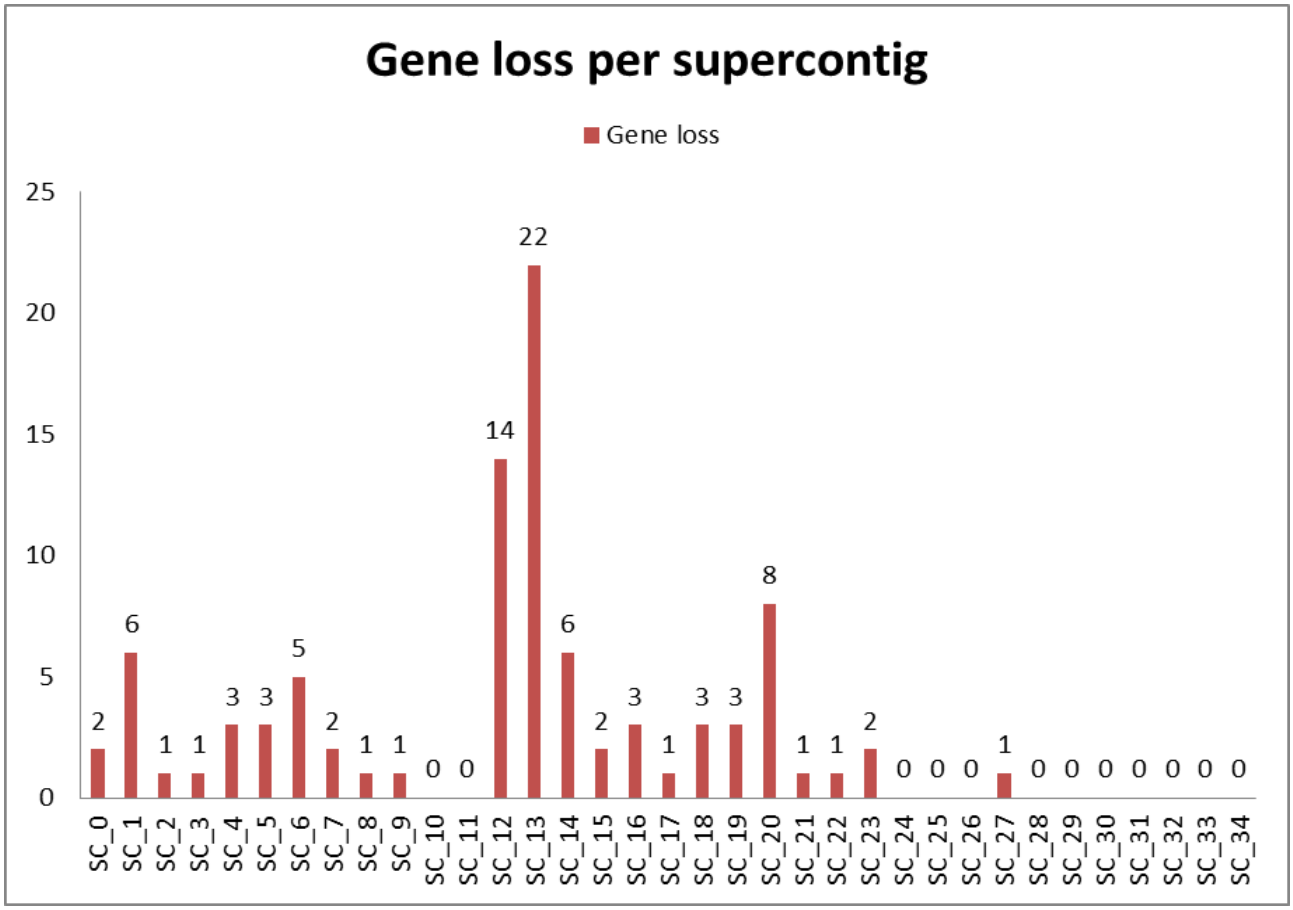

Figure 4.2 Individual gene loss on each supercontig across all isolates

(One count=one gene)

Table 4.3 Gene loss results in isolates 04MGPP021 and 06MGPP041 as listed in Golicz et al. (2015)

(Highlighted Purple-Genes also lost in our data set; 21-04MGPP021; 41-06MGPP041)

		Lost in	
Gene name	**Annotation**	**Isolate 21**	**Isolate 41**
G030170.1	Predicted protein	N	Y
G037480.1	Predicted protein	N	Y
G043120.1	Predicted protein	N	Y
G046740.1	Predicted protein	Y	N
G049660.1	M1 protein (*AvrLm1*)	Y	Y
G049920.1	Predicted protein	N	Y
G049930.1	Predicted protein	N	Y
G049940.1	M6 protein (*AvrLm6*)	N	Y
G076370.1	Predicted protein	Y	N
G070000.1	Histidine triad	N	Y
G079780.1	Predicted protein	Y	Y
G081630.1	Predicted protein	Y	Y
G081640.1	Predicted protein	Y	Y
G082370.1	Predicted protein	Y	N
G082540.1	Glucose-6-phosphate isomerase	Y	Y
G085580.1	Cytochrome p450	Y	Y
G085590.1	Transferase family protein	Y	Y
G085610.1	Predicted protein	Y	Y
G085620.1	Benzoate 4-monooxygenase cytochrome p450	Y	Y
G085630.1	Carboxylesterase family protein	N	Y
G086520.1	DNA repair helicase	N	Y
G086530.1	DNA repair helicase	N	Y
G090230.1	Predicted protein	Y	Y
G098080.1	Serine threonine protein kinase	N	Y
G098090.1	Predicted protein	N	Y
G103490.1	Phosphotransferase family protein	N	Y
G104510.1	Predicted protein	Y	Y

G109260.1	Het domain-containing protein	N	Y
G110260.1	Predicted protein	Y	Y
G113050.1	Tpr domain protein	Y	Y
G113060.1	Predicted protein	Y	Y
G113080.1	Predicted protein	Y	Y
G114280.1	Predicted protein	N	Y
G114380.1	Pyridoxamine phosphate oxidase	N	Y
G114390.1	Mating-type mat 1-2 protein	N	Y
G116710.1	Glycosyltransferase family 34 protein	Y	N
G123850.1	Predicted protein	Y	Y

4.3.3 Gene loss validation

Gene loss predictions for this sample set were validated by detecting known avirulence genes. Preliminary infection studies using inoculum from isolates in this dataset were also conducted to ensure that all isolates were infecting susceptible varieties. The scores are listed in Table 4.4. Table 4.5 provides information on which avirulence genes we expected to be conserved and lost in each sample. Based on the method for conferring virulence being complete deletion, the predictions for *AvrLm1* by this method accurately matched those in Table 4.5 (Data obtained from Dr Angela Van de Wouw, University of Melbourne). *AvrLm6* method of loss in all isolates except for D8, was deletion and the presence or absence of this gene in those isolates was identified accurately. However, *AvrLm6* in D8 is RIP-affected and a small portion of the exon has no reads mapping to it. In this isolate, gene loss did not identify *AvrLm6* since its method of loss was not deletion of the locus. Other known *AvrLm* genes such as *AvrLm2*, *AvrLmJ1, AvrLm3* and *AvrLm11* were not reported by this analysis, as their method of loss of virulence was not gene loss. The first two conferred virulence via point mutations and *AvrLm3* is absent in the reference and thus would not be picked up by this method. *AvrLm11* confers virulence via loss of the minichromosome but was present in all isolates in our dataset. Some lost genes with reads mapping to the gene locus were observed to be mis-mapping of reads.

Table 4.4 Results of preliminary infection studies on 'Westar'
(Westar-Susceptible *B. napus* cultivar containing no resistance genes; Infection scores ranged from 0-9 based on the size of lesions at the site of infection; 0-No lesions 9-Strong lesions)

Name	Infection Score (1-9)
D1	5
D2	8-9
D3	9
D4	5-6
D5	9
D6	8-9
D8	8-9
D9	9
D12	7-9

4.3.3.1 AvrLm6 validation in D8

As per phenotyping studies, *AvrLm6* is virulent in isolate D8. However, it was not reported as lost when SGSGeneLoss was run on this sample. Amplification of this locus using two different primer pairs (Figure 4.3) resulted in the presence of bands, when run on a gel, for two samples of D8 obtained from different sources. Furthermore, replicates of the sample also displayed the same band. This result, combined with phenotyping results leads us to believe that the *AvrLm6* locus in D8 is highly diverged or mutated but nonetheless present in sequence and therefore would not be picked up by the gene loss method. SNPs have been shown to be the cause of virulence in *AvrLm6*.

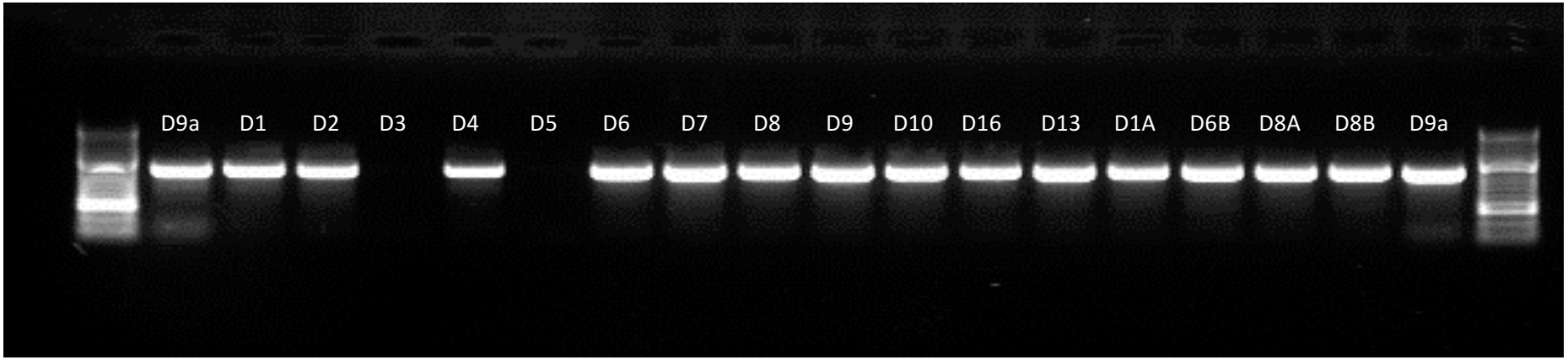

Figure 4.3 PCR using primer pair 2 (Table 4.2) at the *AvrLm6* locus

(a/b-Two replicates of same sample; D1-D9-In lab isolate; D8a/b and D9a/b-Isolates obtained from Prof Martin Barbetti's lab, University of Western Australia; Isolate D3 and D5 do not contain *AvrLm6* as per Table 4.5)

Table 4.5 *AvrLm* gene table deduced from results of infecting *B. napus* cultivars isolates with known *Rlm* gene content, with these isolates

(*AvrLm1-9*- predicted effector genes in *L. maculans*; D1-6, D8, D9 and D12- isolates used in this study; Ref- Reference genome isolate v23.1.3/JN3 (Rouxel et al., 2011); *Avr*- isolate is avirulent towards corresponding resistance genes; *Vir*- isolate is virulent towards corresponding resistance gene and can cause infection; nd- not determined/no data available; Data provided by Dr Angela Van de Wouw, University of Melbourne)

	D1	**D2**	**D3**	**D4**	**D5**	**D6**	**D8**	**D9**	**D12**	**Ref**
AvrLm1	Vir	Vir	Vir	Vir	Avr	Avr	Vir	Vir	Vir	Avr
AvrLm2	Avr	Vir	Vir	Vir	Avr	Vir	Vir	Vir	Vir	Vir
AvrLm3	Vir	Vir	Vir	Vir	Vir	Vir	Vir	Vir	Vir	Vir
AvrLm4	Vir	Vir	Vir	Avr	Avr	Vir	Vir	Vir	Avr	Avr
AvrLm5	Avr	Avr	Avr	Avr	Vir	Avr	Avr	Avr	Avr	Avr
AvrLm6	Avr	Avr	Vir	Avr	Vir	Avr	Vir	Avr	Avr	Avr
AvrLm7	Vir	Vir	Vir	Avr	Avr	Vir	Avr	Avr	Avr	Avr
AvrLm8	Vir	Avr	Vir	Avr	Vir	Avr	nd	nd	nd	Avr
AvrLm9	Avr	Vir	Vir	Vir	Vir	Vir	Vir	Vir	Vir	Vir

4.3.4 Gene loss per isolate

A total of 28 genes were lost uniquely, meaning that a particular gene was lost only in one of the nine isolates analysed (Table 4.6). Overall, each isolate lost between 0-8 genes uniquely, with an average of ~5 unique genes lost per isolate. The isolate D4 lost the most unique genes (eight) and D2 did not lose any genes uniquely. D3 lost the most number of genes compared to the reference; 62, contributing to 16.4% of overall gene loss and D2 lost the least number of total genes compared to the reference at 25. This brought the average gene loss per isolate compared to the reference to 42. As mentioned previously, clustered gene loss was only observed on SC_13 in isolate D4, which lost five genes in a region spanning 10.98 Kb. These five genes were notably part of the sirodesmin biosynthesis cluster, which were otherwise conserved in all other isolates analysed. Amongst uniquely lost genes, where annotations were available, the only common annotations were 'dynamin family protein' (D1 and D5) and 'fad-binding domain containing protein' (two copies lost from D12).

A dendrogram generated using the gene loss data (Figure 4.4) did not display any patterns coinciding with known parameters associated with each sample such as location of sample collection, stubble cultivar or year cultured, as listed in Table 4.6. However, isolates D1, D3, D5 and D9 appeared to group together. These isolates lost the highest number of genes in the sample set, ranging between 45 to 62 genes per isolate (Table 4.6). In the grouping on a separate clade between D2, D4, D6, D8 and D12, D4 appeared to be the most distinct, on a separate branch of the clade (Figure 4.4). The number of genes lost in samples collected in 1987-1988 ranged from 25 to 62 genes lost per sample. Out of nine samples analysed, five were collected from stubble at Mt. Barker, WA in different years. The range of gene loss for these four samples was 31 to 62 genes lost per sample.

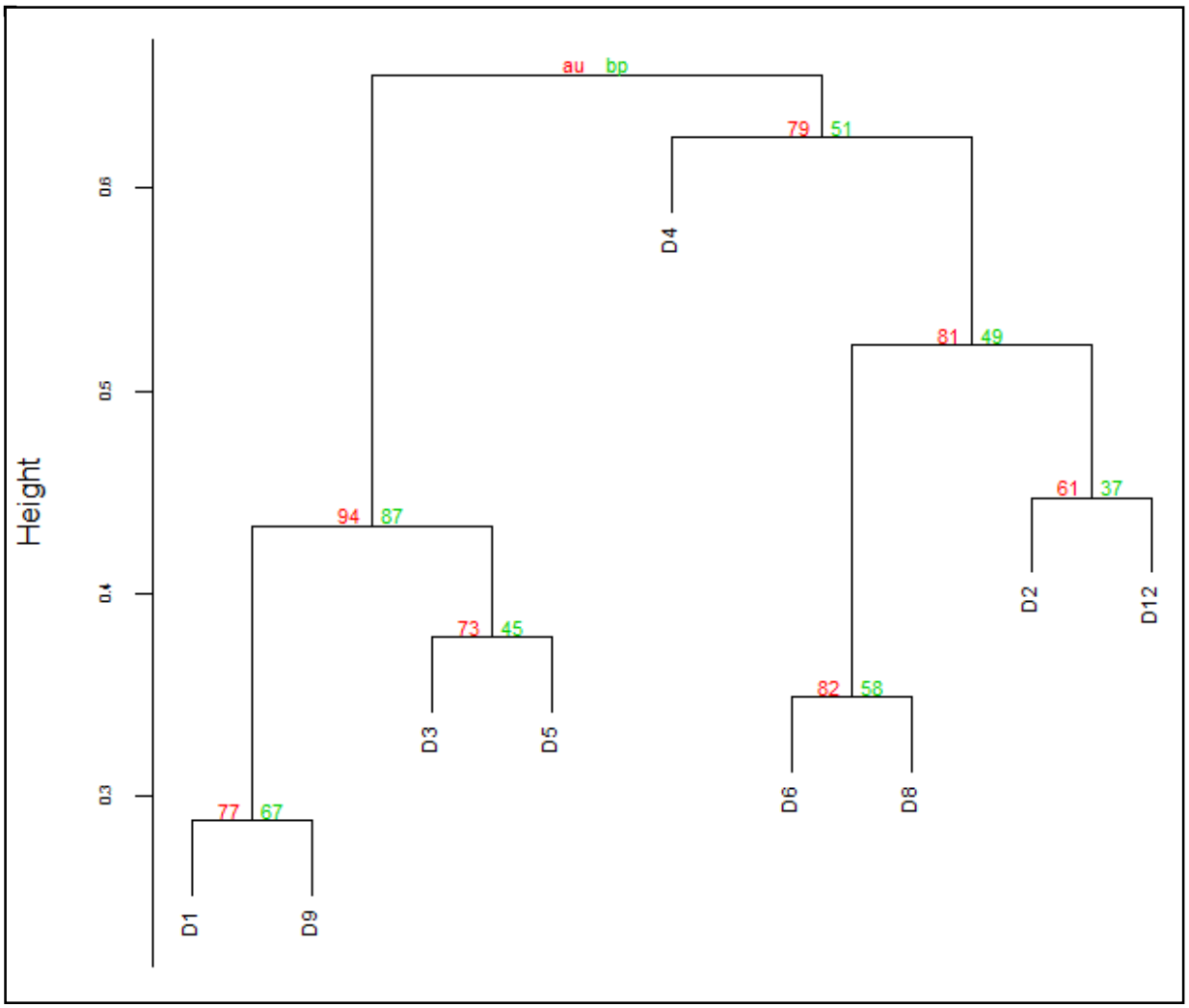

Figure 4.4 Dendrogram showing phylogenetic relationship between samples using results of gene loss

(Bootstrap values: red-Approximately Unbiased (AU) p-values calculated by multiscale bootstrap resampling; green-Bootstrap Probability (BP) p-values calculated by normal bootstrap resampling; AU values >95 % strongly supported by data (Suzuki and Shimodaira, 2006))

.

Table 4.6 Parameters associated with samples used in this study

[WA-Western Australia; Vic-Victoria; SA-South Australia; Ref-Reference genome isolate (Rouxel et al., 2011)]

Isolate Name	**Number of Unique genes lost**	**Number of genes lost**	**Year Cultured**	**Species**	**Stubble cultivar**	**Stubble collection site**	**State/Country**
D1	4	49	1991	*B. napus*	Unknown	Mt Barker	WA
D2	0	25	1988	*B. napus*	Unknown	Streatham	Vic
D3	4	62	1988	*B. napus*	Unknown	Mt Barker	WA
D4	8	41	1988	*B. napus*	Unknown	Millicent	SA
D5	3	45	1988	*B. napus*	Unknown	Penshurst	Vic
D6	1	31	1987	*B. napus*	Unknown	Mt Barker	WA
D8	2	40	2005	*B. napus*	Surpass 501TT	Mt Barker	WA
D9	1	52	2005	*B. napus*	ATR-Beacon	Mt Barker	WA
D12	5	34	2004	*B. napus*	Surpass603CL	Bordertown	SA
Ref/v23.1.3	-	-	1990	-	-	-	Europe

4.3.5 Gene loss on SC_13

SC_13 had the highest number of genes lost across all isolates. It was the only supercontig that displayed clustered gene loss. Isolates D1, D3, D5 and D9 lost twelve genes clustered in a ~27 kbp region (Table 4.7), with a ~5.9 kbp intergenic region (Located between Lema_G081970.1 and Lema_G081980.1) contained within the cluster of lost genes, which had reads mapping to it. A MegaBLAST search in Geneious (Version 6.1.8; Kearse et al., 2012) of this intergenic region in the *L. maculans* genome resulted in 56 sequence hits to different parts of the genome, with >91% pairwise identity and >98% query coverage with an e-value of 0 (Mismatch 1-2; Word size 28), indicating that the reads are mapping to a probable repeat region and are mis-mapped and the entire region is lost in these isolates. In order to gain further confirmation, the sequence was run through RepeatMasker, which identified at least four tandem repeats in the input sequence (Smit et al., 2013-2015). We speculate that this block of genes belongs to a putative secondary metabolite cluster.

Table 4.7 List of genes lost on SC_13 amongst all samples

(Genes thought to be involved in secondary metabolite biosynthesis highlighted purple; Gene loss region in D1, D3, D5 and D9 highlighted orange)

Annotation	**Gene Name**	**D1**	**D2**	**D3**	**D4**	**D5**	**D6**	**D8**	**D9**	**D12**	**Reference**
Predicted protein	Lema_G079780.1	Y	Y	Y	Y	Y	Y	Y	Y	Y	
Predicted protein	Lema_G081630.1	Y	N	Y	Y	Y	N	Y	Y	Y	
Predicted protein	Lema_G081640.1	Y	N	Y	Y	Y	N	Y	Y	Y	
SirH	Lema_G081880.1	N	N	N	Y	N	N	N	N	N	(Gardiner et al., 2004)
Predicted protein	Lema_G081890.1	N	N	N	Y	N	N	N	N	N	
LmUVI-1h	Lema_G081900.1	N	N	N	Y	N	N	N	N	N	
Predicted protein	Lema_G081910.1	N	N	N	Y	N	N	N	N	N	
Polyketide synthase	Lema_G081920.1	N	N	N	Y	N	N	N	N	N	(Gardiner et al., 2004; Rouxel et al., 2011)
LmHDX1	Lema_G081930.1	N	N	N	N	N	N	Y	N	N	(Gardiner et al., 2004)
Aspartate aminotransferase	Lema_G081940.1	Y	N	Y	N	Y	N	Y	Y	N	
Aflatoxin biosynthesis ketoreductase nor-1	Lema_G081950.1	Y	N	Y	N	Y	N	Y	Y	N	
Stress responsive a b barrel domain protein	Lema_G081960.1	Y	N	Y	N	Y	N	Y	Y	N	
Carboxymuconolactone decarboxylase	Lema_G081970.1	Y	N	Y	N	Y	N	N	Y	N	

Salicylate hydroxylase	Lema_G081980.1	Y	N	Y	N	Y	N	N	Y	N	
Predicted protein	Lema_G081990.1	Y	N	Y	N	Y	N	N	Y	N	
Short-chain dehydrogenase reductase sdr	Lema_G082000.1	Y	N	Y	N	Y	N	N	Y	N	
Zinc-binding oxidoreductase	Lema_G082010.1	Y	N	Y	N	Y	N	N	Y	N	
O-methyltransferase	Lema_G082020.1	Y	N	Y	N	Y	N	N	Y	N	
Trihydroxytoluene oxygenase	Lema_G082030.1	Y	N	Y	N	Y	N	N	Y	N	
Predicted protein	Low-read mapping	Y	N	Y	N	Y	N	N	Y	N	
Fatty acid synthase subunit alpha	Lema_G082050.1	Y	N	Y	N	Y	N	N	Y	N	
Fatty acid synthase subunit alpha	Lema_G082100.1	Y	N	Y	N	Y	N	N	Y	N	

4.3.6 Sirodesmin biosynthetic cluster

Five genes (Lema_G081880.1-*SirH*, Lema_G081890.1, Lema_G081900.1- *LmUVI-1h,* Lema_G081910.1 and Lema_G081920.1-Polyketide synthase) that are part of the 68 kbp sirodesmin biosynthesis cluster (Gardiner et al., 2004), were lost uniquely in D4. All other isolates had conserved genes belonging to this cluster.

4.3.7 Functional annotations

Thirty-eight different categories of annotations were predicted using Blast2Go on the gene loss data (Table 4.8). Additional information on these annotations was obtained using Gene Ontology terms associated with the genes. The most number of lost gene annotations were involved in cellular level processes such as oxidation-reduction, glycolysis, gluconeogenesis etc. In addition to these, other genes playing different roles in the infection process were also lost. Some noteworthy annotations of lost genes were *LmUVI-1h* and polyketide synthase in D4, *LmHDX1*-hydroxylase in D8, cutinase on D12, salicylate hydroxylase in D1, D3, D5 and D9, aflatoxin biosynthesis nor-1 ketoreductase in D1, D3, D5, D8 and D9 and beta-glucosidase in D5 (Some lost uniquely highlighted in Table 4.9). The majority of genes lost (58.6%) were annotated as 'predicted protein' whose function is unknown.

Among those genes for which Blast2Go annotations were available, the only similar annotation between uniquely lost genes (Table 4.9) and commonly lost genes (Table 4.10) were different kinds of transferase genes-transferase family proteins, phosphotransferase family protein, and Glycosyltransferase family 34 protein. All commonly lost genes were either predicted proteins or annotated as cytochrome p450, carboxylesterase family proteins or glucose-6-phosphate isomerase which correlate to the most type of genes lost.

Table 4.8 List of available annotations of lost genes in isolates analysed, with position in genome

(Y= lost in that isolate; N=present in that isolate; SC-supercontig; Position-bp location of gene on supercontig; Noteworthy annotations highlighted in pink)

Types of annotations	**D1**	**D2**	**D3**	**D4**	**D5**	**D6**	**D8**	**D9**	**D12**	**SC**	**Position**
Major facilitator superfamily transporter	Y	N	N	N	N	N	N	Y	N	0	763884
Fad binding domain-containing protein	N	N	N	N	N	N	N	N	Y	5	347544
Very large low complexity protein	Y	N	Y	Y	N	N	N	Y	N	5	1055685
Beta-glucosidase	N	N	N	N	Y	N	N	N	N	7	331214
Histidine triad protein	N	N	Y	N	N	N	N	N	N	7	721465
Fungal specific transcription factor domain-containing protein	N	N	N	N	N	N	N	Y	N	8	1427511
Carboxylesterase family protein	Y	Y	Y	Y	Y	Y	Y	Y	Y	12	1193152
Cytochrome p450	Y	Y	Y	Y	Y	Y	Y	Y	Y	12	1145558, 1190933
DNA repair helicase	Y	Y	Y	N	Y	Y	Y	Y	Y	12	1625404, 1628559
Glucose-6-phosphate isomerase	Y	Y	Y	Y	Y	Y	Y	Y	Y	12	222182
Transferase family protein	Y	Y	Y	Y	Y	Y	Y	Y	Y	12	1149272, 1163626
Aflatoxin biosynthesis ketoreductase	Y	N	Y	N	Y	N	Y	Y	N	13	1428596

nor-1											
Aspartate aminotransferase	Y	N	Y	N	Y	N	Y	Y	N	13	1426780
Carboxymuconolactone decarboxylase	Y	N	Y	N	Y	N	N	Y	N	13	1431699
Fatty acid synthase subunit alpha	Y	N	Y	N	Y	N	N	Y	N	13	1451035, 1465862
O-methyltransferase	Y	N	Y	N	Y	N	N	Y	N	13	1446238
Polyketide synthase	N	N	N	Y	N	N	N	N	N	13	1418093
LmHDX1-hydroxylase	N	N	N	N	N	N	Y	N	N	13	1424869
Salicylate hydroxylase	Y	N	Y	N	Y	N	N	Y	N	13	1438204
Short-chain dehydrogenase reductase sdr	Y	N	Y	N	Y	N	N	Y	N	13	1441845
Stress responsive a b barrel domain protein	Y	N	Y	N	Y	N	Y	Y	N	13	1430732
Trihydroxytoluene oxygenase	Y	N	Y	N	Y	N	N	Y	N	13	1448479
LmUVI-1h	N	N	N	Y	N	N	N	N	N	13	1415843
Zinc-binding oxidoreductase	Y	N	Y	N	Y	N	N	Y	N	13	1444412
Calcium calmodulin dependent protein kinase	N	N	Y	N	N	Y	N	Y	N	14	691342
P-loop containing nucleoside triphosphate hydrolase	Y	N	N	N	N	N	N	N	N	14	158218
Serine threonine protein kinase	Y	N	Y	Y	N	N	Y	N	N	14	1291448

Vegetative incompatibility protein het-e-1	Y	N	N	N	N	N	N	N	N	15	555197
Het domain protein	N	Y	N	N	Y	N	Y	N	N	16	513425
Phosphotransferase family protein	N	N	N	Y	N	N	N	N	N	16	289471
Heterokaryon incompatibility protein	Y	N	N	N	N	N	N	Y	N	18	905419
Tetracycline-efflux transporter	N	N	N	Y	N	N	N	N	N	18	31059
Tpr domain protein	Y	N	Y	Y	Y	Y	Y	Y	Y	19	956634
Cutinase	N	N	N	N	N	N	N	N	Y	20	70380
Fad binding domain protein	N	N	N	N	N	N	N	N	Y	20	72098
Fructosamine-3-kinase	N	N	N	N	N	N	N	N	Y	20	69195
Pyridoxamine phosphate oxidase	N	N	Y	Y	N	N	N	Y	N	20	552415
Glycosyltransferase family 34 protein	N	N	N	Y	N	N	N	N	N	21	263332
Dynamin family protein	Y	N	N	N	Y	N	N	N	N	14, 4	152839, 1428187

Table 4.9 Annotation and location information about unique genes lost in the isolates analysed

(Highlighted green-Notable genes lost uniquely)

Isolate	SC	Name	Position	Annotation	Length (aa)	Additional information
D1	14	Lema_G094000.1	152839	Dynamin family protein	811	GTP catabolic process, GTPase activity, GTP binding
D1	14	Lema_G094010.1	157792	Predicted protein	70	-
D1	14	Lema_G094020.1	158218	P-loop containing nucleoside triphosphate hydrolase	3090	-
D1	15	Lema_G091180.1	555197	Vegetative incompatibility protein het-e-1	1076	-
D3	7	Lema_G070000.1	721465	Histidine triad protein	82	Ligase activity, hydrolase activity, translation and nucleotide binding
D3	12	Lema_G086300.1	1400949	Predicted protein	131	-
D3	12	Lema_G086310.1	1402878	Predicted protein	142	-
D3	22	Lema_G119040.1	104915	Predicted protein	34	-
D4	13	Lema_G081880.1	1411646	Predicted protein	407	-
D4	13	Lema_G081890.1	1413647	Predicted protein	35	-
D4	13	Lema_G081900.1	1415843	*LmUVI-1h*	203	-
D4	13	Lema_G081910.1	1417543	Predicted protein	103	-

D4	13	Lema_G081920.1	1418093	Polyketide synthase	2047	Dehydroaustinol biosynthetic process, shamixanthone biosynthetic process, melanin biosynthetic process, violaceol II biosynthetic process, naphtho-gamma-pyrone biosynthetic process, emericellamide biosynthetic process, sterigmatocystin biosynthetic process, arugosin biosynthetic process, transferase activity, transferring acyl groups other than amino-acyl groups, o-orsellinic acid biosynthetic process, austinol biosynthetic process, hydroquinone:oxygen oxidoreductase activity, pigment metabolic process involved in developmental pigmentation, emodin biosynthetic process, asperthecin biosynthetic process,F-9775B biosynthetic process, monodictyphenone biosynthetic process, hydrolase activity, acting on ester bonds, asperfuranone biosynthetic process, violaceol I biosynthetic process,F-9775A biosynthetic process

D4	16	Lema_G103490.1	289471	Phosphotransferase family protein	69	Kinase activity
D4	18	Lema_G106610.1	31059	Tetracycline-efflux transporter	583	Transmembrane transport, integral to membrane, regulation of cellular process
D4	21	Lema_G116710.1	263332	Glycosyltransferase family 34 protein	186	Transferase activity, transferring hexosyl groups, integral to membrane
D5	0	Lema_G004720.1	1574046	Predicted protein	135	-
D5	4	Lema_G043290.1	1428187	Dynamin family protein	757	-
D5	7	Lema_G068460.1	331214	Beta-glucosidase	636	-
D6	20	Lema_G113190.1	201217	Predicted protein	126	-
D8	3	Lema_G037480.1	1850027	Predicted protein	76	-
D8	13	Lema_G081930.1	1424869	*LmHDX1*	476	
D9	8	Lema_G061030.1	1427511	Fungal specific transcription factor domain-containing protein	696	Regulation of transcription from RNA polymerase II promoter, zinc ion binding, DNA binding, nucleus, transcription from RNA polymerase II promoter
D12	5	Lema_G050760.1	347544	Fad binding domain-containing protein	121	Oxidoreductase activity, acting on CH-OH group of donors
D12	20	Lema_G113150.1	57305	Hypothetical protein	475	Heterocyclic compound binding, organic cyclic compound binding, catalytic activity
D12	20	Lema_G113160.1	69195	Fructosamine-3-kinase	260	Phosphatidylinositol metabolic process,

						kinase activity, phosphorylation, ion binding
D12	20	Lema_G113170.1	70380	Cutinase	305	Extracellular region, cutinase activity, cellular carbohydrate catabolic process, lipid catabolic process
D12	20	Lema_G113180.1	72098	Fad binding domain protein	430	Oxidation-reduction process

Table 4.10 Genes lost in common amongst all nine isolates

(Maximum number of common genes lost on SC_12)

Gene name	SC	Location	Annotation
Lema_G019350.1	1	2052355	Predicted protein
Lema_G030170.1	2	2056133	Predicted protein
Lema_G043120.1	4	1393952	Predicted protein
Lema_G046740.1	6	538805	Predicted protein
Lema_G082540.1	12	222182	Glucose-6-phosphate isomerase
Lema_G085580.1	12	1145558	Cytochrome p450
Lema_G085590.1	12	1149272	Transferase family protein
Lema_G085600.1	12	1163626	Transferase family protein
Lema_G085610.1	12	1190615	Predicted protein
Lema_G085620.1	12	1190933	Cytochrome p450
Lema_G085630.1	12	1193152	Carboxylesterase family protein
Lema_G079780.1	13	678403	Predicted protein
Lema_G104510.1	16	667778	Predicted protein
Lema_G110260.1	18	1321392	Predicted protein
Lema_G113080.1	19	1177185	Predicted protein

4.3.8 RNA Seq data

Gene expression data (Lowe et al., 2014) for isolate D5 infecting *B. napus* cv Westar cotyledons was compared against the D5 gene loss profile. Gene expression was analysed in vitro, seven days after inoculation of the fungal sample on a susceptible Brassica cultivar and 14 days after inoculation. Stem canker RNA seq data was also analysed. In vitro, of the 45 lost genes, six genes (13%) reported very low expression (FPKM value between 0-1) whereas the majority, 87% of genes did not have any expression. The gene expression at 7 dpi and 14 dpi ranged from 0 to ~25 FPKM. For both, 47% genes displayed no expression whereas 53% genes had an average expression of 2.08 and 0.25 FPKM respectively. Upon further investigation it was found that the genes reporting these low FPKM values were due to either mis-mapping of reads (genes within multiple gene families and low stringency mapping), or reads belonging to *B. napus* incorrectly mapping to the blackleg genome (as the *B. napus* genome was not used in the initial mapping). Stem canker RNA seq data (Provided by Prof Barbara Howlett's group, University of Melbourne) was also analysed against the D5 loss profile and all genes showed expression values of 0.

4.4 Discussion

With the increase in high-throughput sequencing and availability of reference genomes for several organisms, the gene loss method is a choice tool for analysis of sequence data. This study demonstrates how this method has multi-faceted use to analyse gene content in a set of sequence data. We were able to shed light on new genomic features in individuals of the blackleg population that will assist us in future studies.

Previous work (Chapter 3; Rouxel et al., 2011; Zander et al., 2013; Grandaubert et al., 2014b; Patel et al., 2015) has led us to understand that the Australian blackleg population is highly diverse. Various features of the blackleg genome create an environment conducive to rapid evolution and thereby, genetic diversity. Therefore, we expected to find a varied pattern of gene loss amongst the samples we analysed. However, we also expected that genes essential for survival such as genes encoding multi-drug resistance, actin encoding genes and up-keep of important processes like sexual reproduction would be maintained primarily, whilst genes draining survival resources were expected to be lost.

4.4.1 Gene loss validation

Avirulence genes, such as *AvrLm1* and *AvrLm6,* whose mechanism of conferring virulence is deletion, were successfully detected by the gene loss tool in this sample set. Golicz et al. (2015) also validated these genes in the isolates analysed in their study (Table 4.3). We also found that the *AvrLm6* locus in isolate D8 was subject to RIP mutations and a small portion of the exon had low read-mapping. Fudal et al. (2009) found that *AvrLm6* was subject, not only to deletion of the locus but also RIP mutations and other point mutations, as a means of sequence diversification. Zander (2015) also suggest in their Presence/Absence Variation (PAV) predictions of *AvrLm* genes that *AvrLm6* may be highly diverged in the isolate D8. PCR validation of this gene in D8 confirmed its presence in the genome using PCR and sequence analysis (Figure 4.3). This indicated that although D8 is virulent for *AvrLm6* as per Table 4.5, its method of conferring virulence is not deletion. This would explain why gene loss does not report this gene as lost in D8.

4.4.2 Gene loss statistics

Previous gene loss analysis conducted on two *L. maculans* isolates; 04MGPS021 and 06MGPP041, found that they lost 13 and 27 (0.1% and 0.2% of total predicted genes) genes respectively, when compared to the reference (Table 4.3; Golicz et al., 2015). As seen in Table 4.6, we observed an

average gene loss of ~42 genes per isolate, with no relation to parameters associated with each isolate; such as year of isolation, region of isolation and stubble cultivar isolated from. This is a loss of 0.2-0.5% of total predicted genes in the genome in any given isolate analysed. This suggests that the pathogen maintains the majority of predicted genes as they may play a role in important processes like sexual reproduction and host infection. Occasionally, gaps in assembling the reference genome can also contribute to gene loss. In this scenario, genes that are present in the sample being analysed, may be reported as lost due to gaps in assembly at the location of that gene in the reference genome. This has been exemplified in Mancera et al. (2015) where the ortholog of gene *EFG1* was found in *Candida tropicalis*, where it was previously thought to be absent. Of the 37 genes (Table 4.3) found to be lost in the previous study by Golicz et al. (2015), 33 were also lost in our dataset. Clustered gene loss was noted only on SC_13 (Figure 4.1) likely caused by selection pressure acting in that region of the genome, which is discussed more below. SC_12 had the next highest number of losses per SC (Figure 4.2). SC_12 may host genes involved in non-essential processes, which may explain the high number of genes lost on this supercontig. Genes lost on this SC were primarily lost in pairs, a phenomenon that was previously discovered by Golicz et al. (2015).

The supercontigs SC_24-SC_26 and SC_28-SC_76 that did not show any gene loss, do not form a part of the final nuclear genome. These were reported to either contain unassembled telomeric repeats or telomeric sequences, RIP-affected rDNA copies and unassembled rDNA sequences or be affected by repeat elements containing no protein coding sequence (Rouxel et al. (2011) Supplementary information). Of the three annotated genes on SC_27, one was reported as lost. SC_27 is also reported to contain RIP affected rDNA copies and not a part of the final nuclear genome (Rouxel et al., 2011; Supplementary Information). SC_30 represents mitochondrial DNA which also did not report any lost genes. SC_10 and SC_11 likely contain genes necessary for survival or infection in the pathogen and were therefore conserved. Gene loss was only predicted on supercontigs containing annotated genes in the reference genome. As seen in Table 4.2, sufficient sequence coverage was obtained for each isolate to have confidence in the accuracy of the gene loss prediction for each sample.

4.4.3 SC_13 gene loss

Isolates D1, D3, D5 and D9 displayed clustered gene loss of ~27 kbp on SC_13 (Table 4.7). Based on the annotations associated with the gene loss on this supercontig (Table 4.7), we speculate that these genes belong to a putative secondary metabolite cluster. If a cluster of genes contains a signature secondary metabolite gene such as non-ribosomal peptides (NRPs), polyketide synthases (PKSs) and dimethylallyl tryptophan synthetase (DMAT) homologues, it is probable that those genes are involved in secondary metabolite production (Keller et al., 2005). This cluster of genes can also include backbone genes such as transporters, oxidoreducatases, methylases, esterases, dehydrogenases and acetylases (Keller et al., 2005; Lim et al., 2012). Typically, backbone genes catalyse the first scaffold-forming step in secondary metabolite production. Next, this scaffold is changed by modifying enzymes, usually through a multi-step process, leading to final production (Lim et al., 2012). Fungal PKSs populating secondary metabolite clusters are usually type I PKSs (Keller et al., 2005), as is true for filamentous fungi like blackleg (Elliott et al., 2013). Lema_081920.1 or PKS1 (annotated as polyketide synthase; part of SC_13 cluster) shares 67.5% identity with *pksA*, the PKS gene involved in aflaxtoxin biosynthesis (Genbank accession AAR32704) (Gardiner et al., 2004; Rouxel et al., 2011). The cluster producing sterigmatocystin in *Aspergillus nidulans* encodes some genes similar to those in the cluster on SC_13 (Table 4.7; Brown et al. (1996)). Both alfatoxin and sterigmatocystin are widely known fungal mycotoxins produced by the *Aspergillus* species. The biosynthesis of both are encoded in 75 kb (Yu et al., 2004) and 60 kb (Brown et al., 1996) secondary metabolite clusters. Detailed functionality of this putative secondary metabolite cluster can be elucidated once further functional studies are conducted. This could facilitate characterisation of other genes located around the cluster that could play a role in potential secondary metabolite production.

Secondary metabolite clusters are so termed due to their dispensability and their redundant role in survival of the organism. Furthermore, these clusters are usually activated upon stimulus by a particular environmental cue, evolving as a response to predation and harsh environments (Lim et al., 2012). As is the case here, these clusters are conserved in some isolates we analysed but lost in others. It has been suggested that fungi adopt secondary metabolite clusters via horizontal transfer (Keller et al., 2005; Grandaubert et al., 2014b). However, certain other clusters like that in the fungus *Gibberella fujikuroi* producing gibberellin show that some clusters are partially conserved within the genome (Keller et al., 2005).

A BLAST search of the ~5.9 kb intergenic region, contained within the clustered gene loss, in the *L. maculans* genome showed that this region is most likely repetitive. Rouxel et al. (2011) state that the blackleg genome is riddled with mosaics of truncated transposable elements (TEs). The genome is made up of 32.5% TE content, highest among the pathogen's closest relatives (Grandaubert et al., 2014b). Multiple high-identity BLAST hits indicate that reads mapping to that region could just as easily map to another region of the genome, depending upon the mapping parameters. This leads us to conclude that SC_13 gene loss is a true deletion of a large locus signifying high selection pressure acting in that region. Results from RepeatMasker (Smit et al., 2013-2015) identifying four tandem repeats in the region, gave our conclusion further confidence. Another probability could be that the intergenic region with reads mapping could be a mis-assembly and be located in another part of the genome.

4.4.3.1 Sirodesmin biosynthesis cluster

Sirodesmin is a fungal phytotoxin that causes yellow lesions on infected plants and is non-host specific (Rouxel et al., 1988). Mutant *L. maculans* strains lacking sirodesmin production abilities displayed smaller lesions on infected plants and a reduced ability to colonise stems (Elliott et al., 2007); however, it is known to be suppressed by the phytoalexin brassinin of canola (Pedras and Taylor, 1993). The cluster of genes responsible for sirodesmin biosynthesis is also located on SC_13. This 68 kbp cluster contains 23 genes, of which 18 are thought to be involved in producing sirodesmin (Genbank accession AY553235). PKS1, which is a part of the putative cluster on SC_13, was characterised as a neighbouring gene but not thought to be involved in synthesis (Gardiner et al., 2004). *LmUVI-1h* (Lema_G081900.1) and *LmHDX1* (Lema_G081930.1) were also thought to not be involved in the process. The former is expressed under UV light and during fungal appresorium formation and the latter functions as a hydroxylase.

Predictably, genes involved in sirodesmin production were conserved in all isolates except D4. Even within D4 gene loss, *SirH* (Lema_G081880.1) is the only gene directly involved in sirodesmin production. It is an acetyl transferase which plays a role in the final step of sirodesmin production. One might speculate that loss of *SirH* will affect the production of sirodesmin in D4 and subsequently, the ability of the isolate to colonise stems. Previous studies have also found reduced expression of this gene leading to the presence of high amounts of "deactyl derivative of sirodesmin PL" (Gardiner et al., 2004). However, the biological importance of this still needs to be deciphered.

4.4.4 Common lost genes and SC_12 gene loss

Fifteen genes were lost in common amongst isolates, as compared to the reference genome (Table 4.10). Available annotations for these genes include cytochrome p450s, transferase family protein (TFPs) genes, carboxylesterase family protein (CFPs) genes and glucose-6-phosphate isomerase (GPIs) genes. Notably, these four types of genes were all located on SC_12. Their characteristic of being commonly lost attributes the highest number of genes lost to SC_12.

Golicz et al. (2015) also reported four of these genes to be lost commonly in their analysis (Table 4.3). Of these, cytochrome p450 gene (Lema_G085620.1) was the available annotation lost in both isolates analysed in that study. In our results, two copies of cytochrome p450 were found to be lost in every isolate analysed (Lema_G085580.1 and Lema_G085620.1). According to the cytochrome p450 database (Park et al., 2008), the *L. maculans* genome contains 62 distinct cytochrome p450s that represent 0.5% of overall predicted genes. Both these cytochrome p450 genes lost are predicted to play a role in the oxidation-reduction process. GPI genes are involved in the processes of glycolysis and gluconeogenesis. These genes play a catabolic role in the interconversion of D-glucose-6-phosphate and D-fructose-6-phosphate (Ospina-Giraldo and Jones, 2003). Two copies of genes belonging to the transferase family were also lost commonly. Previous studies have found that genes belonging to certain classes are more volatile and the pathogen tolerates their losses and duplication better than genes involved in essential cellular processes eg. Related to growth. They also note that such losses and duplications may even afford a beneficial advantage to the pathogen to adapt to diverse ecological niches (Wapinski et al., 2007). We speculate that commonly lost genes (Table 4.10) belong to this volatile group of genes, whose absence in the genome is tolerable by the pathogen. Furthermore, there may be possible genetic redundancy, as in the case of cytochrome p450, where loss of these genes may not affect functionality in the genome.

4.4.5 Gene loss per isolate

Our results show that gene maintenance and loss amongst the samples we analysed, varied in each isolate (Table 4.6), with some commonalities (Table 4.10). Several reasons contribute collectively to explain why each isolate differs in the genes it maintains and loses.

External environmental factors

The gene loss pattern observed within each isolate shows that there is diversity in conserved versus disposable genes even within species. As seen in Table 4.8, annotations for genes lost uniquely are largely different to each other. As a majority, no one type of gene is lost preferentially, signifying redundancy for single losses in the genome. The genome of each isolate can be shaped based on what environment it was isolated from; factors such as temperature, rainfall or cultivars grown in the region of isolation, period of growth may all cause varied diversifying effects on the genome (Table 4.6). This could explain the variable number of genes lost per isolate and even the number of genes lost uniquely differs within this data set, with each isolate being subject to a different combination of selection pressures from these external factors. We were unable to directly compare phylogenetic analyses between Chapter 3 and Figure 4.4 due to the difference in sample size and type of data input. However, phylogenetic analyses in Chapter 3 grouped isolates D2, D6 and D8 is closely related sub-branches and similarly D3 and D9. Isolates D4 and D5 were also similarly grouped. However, in order to successfully validate the effect of environmental factors and the relationship between isolates, gene loss analysis in a larger sample size, sampled from different growing regions within Australia must be conducted. The effects caused by a range of different climate variables must also be considered. Similar conclusions were also derived from SNP studies conducted on these, amongst other samples of the pathogen in Australia (Patel et al., 2015; Chapter 3).

Overcoming host resistance

The basis of plant-biotic interaction is formed by the following effects, either of which can be triggered when a pathogen infects a host plant. Effector-triggered susceptibility (ETS) is defined as the changes effected by the pathogen in the plant to cause infection (Win et al., 2012). These may include but are not limited to, suppressing the plant's immune system and changing its physiology. Effector-triggered immunity (ETI) occurs when the host plant recognises the pathogen effector due to the presence of a mediating immunoreceptor in the plant (Win et al., 2012). In such a scenario, these pathogens rapidly evolve to overcome the effects of ETI and maintain their virulent nature. Locus deletion, RIP mutations, loss-of-function SNPs, gene gains and losses are all different

processes occurring in rapidly evolving species to bring about change in the genome (Gladieux et al., 2014). This sort of adaptive divergence is consistent with the permanent arms race the pathogen is involved in with its host. Previous studies have established that *L. maculans* possesses a high evolutionary potential and rapidly evolves to overcome host resistance (Chapter 3; Barrins et al., 2004; Rouxel et al., 2011; Travadon et al., 2011; Dilmaghani et al., 2012; Zander et al., 2013; Grandaubert et al., 2014b; Patel et al., 2015). Gene loss via deletion is an effective method of adaptation employed by several fungal pathogens like *Magnaporthe oryzae* (Couch et al., 2005), *Blumeria graminis* (Wicker et al., 2013), including *L. maculans* (Gout et al., 2006b; Fudal et al., 2007). We hypothesised that one of the major causes for gene loss in this pathogen would be deletion in known effector genes and other pathogenicity related or unknown effectors, to overcome host recognition. This has been validated in previous studies of gene loss in *L. maculans* (Golicz et al., 2015) and this study (See section 4.3.4). We have detected the loss of known *AvrLm* genes in this sample set, validating the theory that the pathogen loses 'compromised' avirulence genes through the method of gene loss. Furthermore, this method of evolution also enables the pathogen to dispose of any unnecessary energy-consuming enzyme-production mechanisms (Gladieux et al., 2014) as we speculate, is the case for gene loss in this sample set.

Effect of transposable elements on genome architecture

The *L. maculans* genome comprises a high density of transposable elements (TE) (Rouxel et al., 2011). Other studies have also described their role in shaping genomic architecture of this pathogen, affording this pathogenic genome its plastic nature (Grandaubert et al., 2014b). TE activity is known to cause a range of subsequent variations like deletions, insertions, translocations and inversions in the genomes affected by them (Ayarpadikannan and Kim, 2014). At the DNA level, TEs are able to modify the genes close to or at their site of insertion. This may lead to changes in protein products, expression of the gene or alternative splicing (Hua-Van et al., 2011). These modifying effects may also extend over several kb if a particular TE is accompanied by silencers. If these modifying effects afford a fitness cost to the genome, the organism may also lose affected genes over the course of time due to their deleterious nature. Other effect of TEs include recombination between two copies of TEs, altering or silencing gene expression due to insertion in coding sequence, deletions and chromosomal effects caused due to TE-induced recombination (Grandaubert et al., 2014a).

As per the theory of 'Clade selection' (Williams, 1992), we understand that species whose genomes are highly plastic and flexible, can evolve at-speed with their host counterparts, to combat their

immune systems. This ability assists in quick adaptation during co-evolution and maintaining the virulent nature of the pathogen (Raffaele and Kamoun, 2012). In *L. maculans*, it is known that disease-causing effector genes are usually present in a part of the genome enriched by TEs in order to enable rapid sequence diversification and adaptation to host pressures (Grandaubert et al., 2014b). This diversification is assisted by processes like RIP mutations that cause mutations in unfavourable genes to the point of eventual deletion. On the other hand, RIP can also help in mutating effector genes to create novel sequences (Grandaubert et al., 2014a).

Other studies, such as that in *M. oryzae* (Chuma et al., 2011), have found *Avr-Pita*, the pathogenicity gene for this species, is translocated multiple times and deleted in some cases in the genome due to it being closely linked to a retrotransposon. Grandaubert et al. (2014b) describe TEs as 'drivers of gene innovation'. In the case of *L. maculans*, effector genes are displaced to AT-rich TE-affected plastic regions of the genome.

RNA seq data

We were able to correlate the results of gene loss in isolate D5 to RNA seq expression data available for lost genes in vitro, 7 dpi and 14 dpi. The majority of genes lost had no expression in vitro, as expected. Some genes had FPKM expression values between 0-1 indicating very low expression and others an average of 2.08 and 2.54 FPKM expression values at 7 dpi and 14 dpi. FPKM is the ratio of RNA-seq reads that align to a gene. The FPKM value given to any particular gene's expression is affected by the length of the coding sequence and the number of reads that align. For genes with these expression values, the stringency of calculation may allow for inaccurate mapping of RNA-seq reads to deleted gene loci. The RNA seq cotyledon data was obtained from Lowe et al. (2014) which did not use mapping parameters to match the stringency of only obtaining expression for genes truly present, indicating that these reads are mis-mapping which was confirmed through further, more stringent mapping analysis. Another likely explanation for the very low levels of expression observed is that the RNA seq results may be reporting expressions of unannotated genes that are closely related to lost genes and appending the results for the unannotated genes to the lost genes. Therefore, genes that are closely related to lost genes are actually being expressed but the results depict this expression to be that of lost genes, through mis-mapping of reads. Some lost genes in D5 belong to families of related genes eg cytochrome p450. Given that some of the genes are only expressed upon infection, they may also be reads from plant genes, which make up the majority of the RNA-Seq reads upon infection. The deleted status of these genes is supported by stem-canker RNA-seq data where none of these genes are expressed.

4.4.6 Functional annotations and unique genes lost

Excluding known avirulence genes, thirty-eight genes with an annotation were analysed in the gene loss data set. The majority of genes lost are thought to belong to the volatile group of genes whose loss does not afford a fitness cost to the pathogen. Some noteworthy annotations are discussed below.

Beta-glucosidase (Lema_G068460.1)

This gene was lost uniquely in isolate D5 and is located on SC_7. Of 132 identified families of β-glucosidases, family 3 comprises this enzyme for bacteria and fungi (Tiwari et al., 2013). They are a part of the cellulase enzyme complex along with exoglucanase and endoglucanase that cleave glycosidic linkages (Tiwari et al., 2013; Singhania et al., 2013). They play a role in hydrolysing cellobiose and short chain oligosaccharides to glucose where β-glucosidase enzymatic activity reduces as the glucose chain lengthens (Bhatia et al., 2002). Previous experiments on the *bgl1* gene in *Trichoderma reesei* that encodes for an extracellular β-glucosidase, found that deleting the gene in the mutant strain did not affect growth of the strain on cellulose (Fowler and Brown, 1992). The study found that although the absence of *bgl1* caused delayed induction of cellulase enzymes, it had no effect on growth or mRNA production abilities of the cell. However, other gene disruption studies on β-glucosidase genes produced different results (Mishra et al., 1989; Strauss and Kubicek, 1990), leading us to believe that specific functional analysis of this gene in *L. maculans* is required. Preliminary infection studies using D5 did not produce any notable delays in infection or growth of the pathogen in planta (Table 4.4). RNA seq data obtained from Lowe et al. (2014) did not have any reads mapping to the gene locus in vitro, 7dpi or 14dpi, indicating that the gene is truly lost.

Aflatoxin biosynthesis ketoreductase nor-1 (Lema_G081950.1)

This gene was lost in the isolates D1, D3, D5, D9 and D8. The gene is located on SC_13 and is part of the clustered gene loss observed in the first four isolates (Section 4.3.2.2). As discussed earlier, this gene is proposed to be a part of a novel secondary metabolite cluster. Previous functional studies of other genes encoding similar protein products to this gene found that its product plays a catalytic role in converting norsolorinic acid to averantin during the biosynthesis of aflatoxin (Zhou and Linz, 1999). Being a part of a cluster synthesising a secondary metabolite, a varied pattern of conservation vs loss is expected in each sample (See section 4.4.1.1).

LmUVI-1h (Lema_G081900.1), Polyketide synthase (Lema_G081920.1)

The ultraviolet inducible gene *LmUVI-1h* is described in Gardiner et al. (2004) as a gene present close to the sirodesmin biosynthesis cluster, but not thought to be a part of the cluster (Table 4.7). It shares similarity with the *Uvi-1* gene (Genbank accession number BAA96294) of *Bipolaris oryzae*. Under UV exposure, the transcript for this gene accumulates and does so during appresorium formation (Gardiner et al., 2004). This gene is lost uniquely in isolate D4.

Polyketide synthase or PKS1 was uniquely lost in isolate D8. This gene was previously identified by Rouxel et al. (2011) who stated that it shared 67.5% similarity with the PksA gene of *Aspergillus flavus.* The authors also identified eleven other polyketide synthase genes in the blackleg genome, which play a role in secondary metabolite biosynthesis.

Secondary metabolites are enzymes that assist the pathogen in the infection process. It is possible that the products of other functional PKSs may compensate for the loss of this gene in D8. On the other hand, the blackleg genome contains only one copy of the *LmUVI-1h* gene, the loss of which may affect the pathogen. This can be backed-up by our preliminary infection data (Table 4.4) where the D4 infection score on susceptible cultivars is lower than D8. Gardiner et al. (2004) also found that PKS1 and *LmUVI-1h* expressions levels were strongly correlated. However, further gene characterisation involving gene knockouts is needed to pin-point if the absence of *LmUVI-1h* in D4 is detrimental to its infection prowess. Gene complementation studies can also be conducted by introducing *LmUVI-1h* in isolate D4.

LmHDX1 (Lema_G081930.1) and salicylate hydroxylase (Lema_G081980.1)

LmHDX1 is a hydroxylase gene bordering the sirodesmin cluster which shares homology with the *Pseudomonas putida* hydroxylase gene (Gardiner et al., 2004). This gene is located on SC_13 and is lost only in D8. Salicylate hydroxylases convert salicylic acid to catechol (Ambrose et al., 2015). Lema_G081980.1 is lost in isolates D1 D3, D5 and D9 (Table 4.5; Table 4.3). Salicylic acid is known to play an important role in plant defence signalling. Clay and Schardl (2002) found that fungal salicylate hydroxylase, *NahG* breaks down plant salicylic acid to catechol and the reduced amounts of salicylic acid enables the pathogen *P. putida* to cause infection. However, results detailing the function of the same gene in *Ustilago maydis* (Rabe et al., 2013) and in *Epichloe festucae* (Ambrose et al., 2015) opposed this. Both studies stated that the absence of salicylate hydroxylase did not affect pathogen virulence. This leads us to understand that salicylate hydroxylase gene is disposable in *L. maculans*.

Cutinase (Lema_G113170.1)

Cutinase functions to hydrolyse plant polymer bonds (such as in the plant cuticle) including low molecular weight soluble esters and insoluble triacylglycerols (Chen et al., 2013). Located on SC_20 this gene is lost only in the isolate D12. Recent work on the gene suggests that a pathogen may contain several cutinases, each playing a role at different points in the life cycle such as host infection or saprophytic growth (Skamnioti et al., 2008a; Skamnioti et al., 2008b). This study also characterised five fungal species (*Aspergillus nidulans, Botrytis cinerea, Fusarium graminearum, Magnaporthe grisea* and *Neurospora crassa*) for their cutinase gene content, which ranged from 3 to 17 putative cutinases in each genome. Different species gained or lost this gene at different rates. We speculate that *L. maculans* may contain other cutinase genes that are yet to be characterised. Based on the findings of the previous study, we expected these genes to be conserved in *L. maculans* which is highlighted in our data. This gene is only lost in D12 but our data does not suggest that loss of this gene may affect the infection ability of this isolate.

4.5 Conclusion

Over the course of this analysis, we gained an understanding of important genes in the blackleg genome through initial functional characterisation and discovered the difference in gene conservation amongst isolates. We were also able to highlight D4 as an isolate of interest because of its particular pattern of loss. SGSGeneLoss helped us highlight regions under selection pressure in the genome. Blackleg maintains a large proportion of its predicted genes but can easily lose a small portion of those for a variety of reasons. This may be functional redundancy, effect of RIP mutations and TEs or lightening the energy load these genes impose on the pathogen. Analysing gene loss data in *L. maculans* provided us with new genomic insights and several intriguing results that will be explored further.

5 Avirulence gene prediction and genomics of *Leptospheria maculans* using three approaches

5.1 Introduction

Whole genome sequencing and analysis of genomic sequence data has proved to be very useful in analysing genomic particulars of an organism. The availability of large genomic data-sets has led to the development of several bioinformatics tools that assist the user in understanding the biology behind sequencing details. SGSGeneLoss (Golicz et al., 2015), SGSautoSNP (Lorenc et al., 2012) and Present/Absent K-mer Analysis Pipeline (PAKAP) (Zander, 2015) prediction are three different bioinformatics tools that highlight various genomic features of a sample in three different manners. Each method is useful in identifying one aspect of information of interest to the user. However, utilising all three methods at once can provide us with a robust understanding of the sample under scrutiny. Detailed here is a combined use of all three approaches for the identification of novel effector genes that play a pivotal role in causing fungal infection.

5.1.1 Avirulence gene prediction

According to the gene-for-gene interaction or effector triggered immunity, certain pathogen avirulence genes are recognised by their corresponding host resistance genes (Ansan-Melayah et al., 1998; Dangl and Jones, 2001). The pathogen employs several techniques to bring about changes in the genome to overcome this recognition as depicted in Figure 5.1. As has been shown in several cases, this may occur by way of frameshift mutations, non-synonymous point mutations, transposon insertion and in some cases, deletion of the gene.

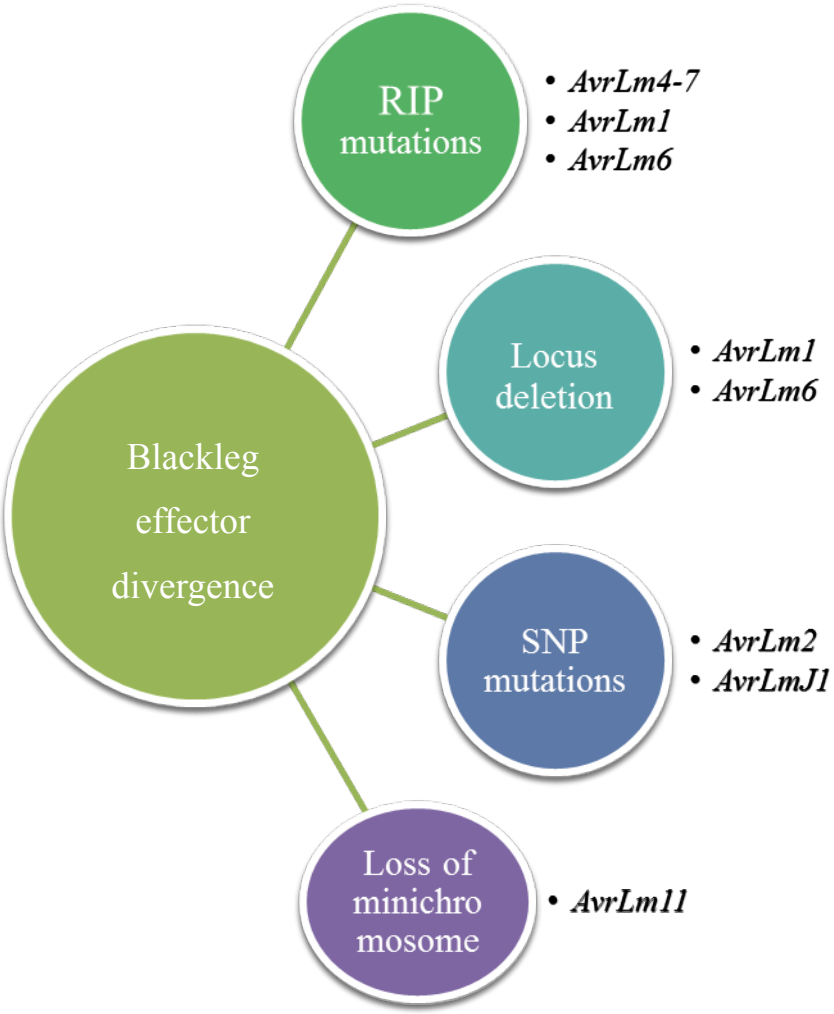

Figure 5.1 Pathways of divergence in *L. maculans*

Identification of novel effector genes is usually challenging due to the low degree of conservation amongst known effectors (Stergiopoulos and de Wit, 2009). This is largely attributed to host specialisation and their ability to evolve rapidly (Sperschneider et al., 2015). BLAST searches in sequenced genomes will only prove effective when known effectors share sequence homology to putative effectors (De Wit et al., 2009). Due to this, effector prediction can only be based on general criteria several of which are discussed in Section 1.4 of chapter 1. Some of these are the common attribute of effectors being cysteine rich or small secreted proteins. (Sperschneider et al., 2015) have compiled a comparative analysis of cysteine content of effectors of various fungi. This shows that previously used thresholds to filter out putative effectors still do not encompass all known effectors based on their cysteine content. This means that although this may be a highlighting feature for several effectors, it does not hold true for all effectors (Sperschneider et al., 2015). Prior criteria for

filtering out candidates recommended by Saunders et al. (2012) and Ma et al. (2010) also fail to take into account all effectors.

It is recommended that one adopts a multi-faceted approach when mining for putative effectors in genomic sequence data. Sperschneider et al. (2015) suggest that other attributes such as information about taxonomy, expression data and evidence of diversifying selection, should also be taken into account. Integrating this with previous methods will establish an unbiased approach for future effector predictions.

To date, seven avirulence genes of *L. maculans* have been characterised. Initially, map-based cloning was used to identify *AvrLm1* and *AvrLm6* (Gout et al., 2006b; Fudal et al., 2007), which are clustered on SC_6, along with *AvrLm2* (Ghanbarnia et al., 2014), in a region affected by RIP mutations. In isolates virulent on cultivars containing *Rlm1* and *Rlm6*, the loci containing *AvrLm1* and *AvrLm6* respectively, were deleted (Gout et al., 2006b; Fudal et al., 2007; Van de Wouw et al., 2010; Zander et al., 2013). Fudal et al. (2009) and Van de Wouw et al. (2010) also found evidence of RIP mutations at the *AvrLm6* locus. *AvrLm2* was diversified by the effect of three SNPs in its coding sequence (Ghanbarnia et al., 2014). In the case of *AvrLm4-7*, *Rlm4* virulence is conferred via a single amino acid substitution in *AvrLm4,* whereas *Rlm7* virulence is conferred through RIP mutations and locus deletion of *Avr7* (Daverdin et al., 2012; Parlange et al., 2009; Van de Wouw and Howlett, 2012). *AvrLm11* is lost via loss of conditionally dispensable chromosome to avoid host detection. Lastly, *AvrLmJ1*, avirulent towards *Brassica juncea* cultivars, becomes virulent via a single mutation that causes a premature stop codon within the coding sequence (Van de Wouw et al., 2014). *AvrLm3*, the most recently discovered avirulence gene was found on SC_12 and its phenotype was found to be masked by the presence of a functional *AvrLm4-7* allele (Plissonneau et al., 2015).

5.1.2 Three approaches for *L. maculans* avirulence gene identification

Due to the difficulty in identifying effectors through simple genomic approaches, it is important that the lifestyle and infection methodology of the particular organism being studied be taken into consideration. The combination of prior knowledge about known effectors and behaviour patterns of the pathogen with bioinformatics tools designed keeping these traits in mind is a promising approach to identifying novel effectors. The following three methods were applied to *L. maculans* avirulence gene identification.

1. SGSGeneLoss predicts genes that are lost or deleted in one sample versus another by calculating the fraction the exon on an annotated gene on the reference genome, is covered by sequence reads. However, this method does require a reference genome for prediction (Golicz et al., 2015). This tool helps characterise effectors based on the knowledge that they may be deleted to overcome host recognition.
2. PAKAP or Presence/Absence variation (PAV) prediction identifies unique reads present only in one sample compared to others and does not require a reference genome (Zander, 2015). This method helps identify those effectors that are not present in the reference genome without the need to sequence and assemble whole genomes. Where a reference genome is available, it also highlights genetic uniqueness amongst the sample set, bypassing the need for utilising computationally challenging analyses.
3. SGSautoSNP predicts single nucleotide point mutations between the samples provided as input (Lorenc et al., 2012). Zander et al. (2013) and Patel et al. (2015) applied this tool to two different sample sets. The former highlighted the efficacy of this tool in identifying differences between two blackleg isolates without the need of a reference genome whereas the latter used the SNP resource generated from this tool for population genetic analyses of the pathogen.

Predictions from these bioinformatics tools supported by RNA seq expression data (Lowe et al., 2014), provided by Prof Barbara Howlett's group, (University of Melbourne) has been used for effector gene identification detailed in this chapter.

5.2 Material and methods

5.2.1 Sample growth and tissue harvest

The isolates D1, D2, D3, D4, D5, D6, D8, D9 and D12 (See Chapter 3, Table 3.1 for isolate details) were grown and tissue was harvested as per the method detailed in Chapter 2, Section 2.1 and Section 2.3.

5.2.2 DNA extraction, quantification and quality analysis

DNA extraction of samples was conducted using the method listed in Chapter 2, Section 2.4. DNA was quantified using the Qubit Fluorometer (Life Technologies) as per the manufacturer's instructions (Details in Chapter 2, Section 2.5.2). After quantification, DNA samples were run on a 1% agarose gel to ascertain quality (Chapter 2, Section 2.5.1).

5.2.3 Library preparation, sequencing and reference genome

See Section 4.2.5, chapter 4 for details of library preparation and sequencing. The reference genome for *L. maculans* was obtained from Rouxel et al. (2011), strain v23.1.3.

5.2.4 Present/Absent K-mer Analysis Pipeline (PAKAP) prediction

PAKAP prediction was developed by Dr Kenneth Chan. This method highlights unique reads present in one isolate as compared to others in the set (e.g. unique in D1 and not D2, D1 vs D2). It is a K-mer based approach; K-mers being smaller subsets of whole sequence reads. The method begins by identifying an optimal K-mer size for each sample in the input. Next, it highlights K-mers that are unique to each sample and finally pulls out whole sequence reads corresponding to each unique K-mer. Details of the working of this method can be obtained from Zander (2015).

5.2.5 SGSautoSNP

Default parameters in SGSautoSNP (Lorenc et al., 2012) were employed for SNP prediction between samples. Refer to Zander et al. (2013) for details.

5.2.6 SGSGeneLoss

SGSGeneLoss prediction was conducted on samples as per the method listed in Chapter 4, Section 4.2.7.

5.2.7 SOAP mapping

DNA sequence reads for samples were mapped using SOAPaligner 2 (Li et al., 2009). The parameters were set to –r 0 in order to ensure that each read maps to the genome only once. –M 2 allowed for the presence of two mismatches while mapping.

5.2.8 PROtein Variation Effect Analyser (PROVEAN) prediction

Effect of SNP mutations was predicted using PROVEAN (Choi et al., 2012). Amino acid residues with their corresponding variants after SNP mutations were inputted into the PROVEAN web server. The cut-off delta score for deleterious vs neutral mutation was set to default at -2.5 (Below -2.5 mutations are deleterious and above are neutral).

5.2.9 Gene annotation

Gene annotation for SSPs was performed using Blast2Go (Conesa et al., 2005) as detailed in Section 4.2.8, Chapter 4. However, the database used for BLAST analysis was ‘b2g_sep15’.

5.2.10 Linkage disequilibrium analysis

LD analysis was conducted as per the method detailed in Section 3.2.4, Chapter 3.

5.3 Results

The pooling of PAV, gene loss and SNP data resources led to the identification of four effector gene candidates. We were also able to perform preliminary genomics that validated previous results and solidified the understanding that the genomic environment of each isolate differs greatly from one another.

5.3.1 Candidate effector gene identification using PAVs, SNPs and gene loss

As per previous studies, *AvrLm* genes in blackleg can be diverged in the pathogen through several pathways (Figure 5.1). These include point mutations and locus deletions. Keeping in mind these findings, we discovered candidate effector genes on the basis of Gene Loss, PAV and SNP analysis.

5.3.1.1 Pipeline

We developed a basic pipeline to identify novel effectors in *L. maculans*. Figure 5.2 details the steps of the pipeline to effector discovery.

Step 1

We obtained sequence reads for isolates D1, D2, D3, D4, D5, D6, D8, D9 and D12, details of which can be obtained from Table 4.1 (Chapter 4). The optimal amount of coverage required for accurate predictions was at least 10x for each sample.

Step 2

SGSGeneLoss, SGSautoSNP and PAKAP prediction was conducted on each sample.

Step 3

Each set of three types of data was combed to identify potential effectors. The gene loss method was also used to identify novel effector gene candidates. As described previously, deletion of the locus containing the effector gene is one of the methods employed by the pathogen to evade host detection. These results were combined with analysing the AT content of the region comprising the gene, location of the gene, length of gene and proposed cysteine content of the predicted protein.

For PAKAP predictions, the list of PAV reads mapping to known gene annotations was formulated. This was then analysed to highlight potential reads mapping according to patterns for *AvrLm* genes in the differential data.

SNPs were identified in genes of interest present in the reference genome, highlighted by PAVs and analysed for any role they may play in sequence divergence. Signs of RIP mutations acting on a particular region were also searched for.

For potential candidates without known gene annotations, that sequence of interest was also run through SignalP (Petersen et al., 2011) to identify the presence of probable signal peptide cleavage sites. This was followed by Blast2Go analysis (Conesa et al., 2005) on each effector to obtain more information about their protein products.

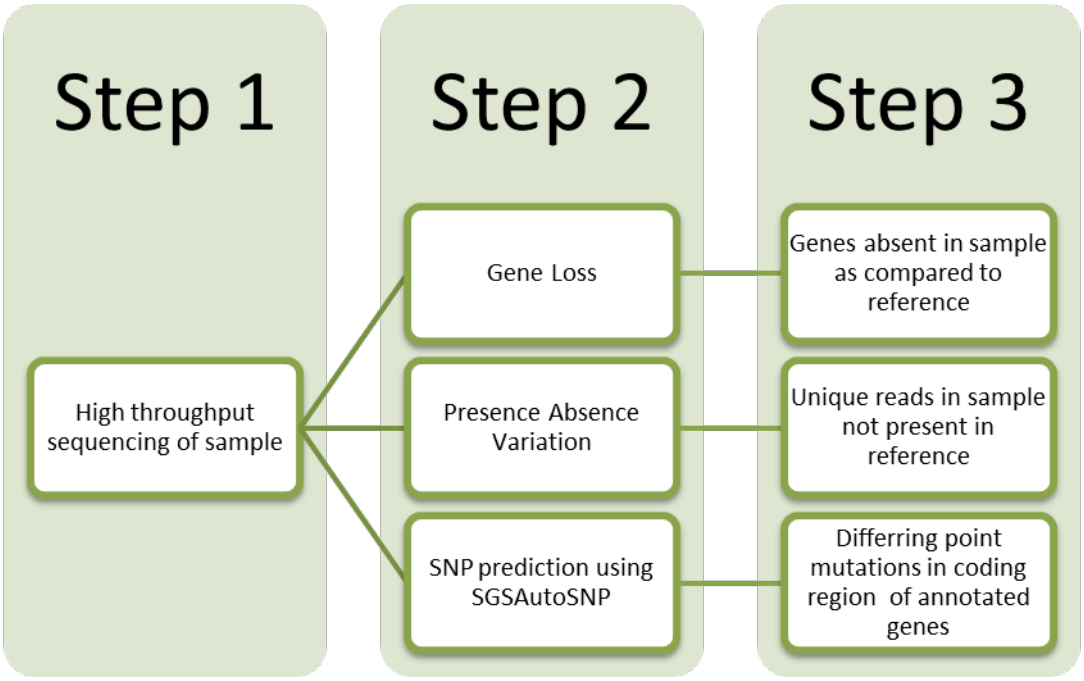

Figure 5.2 Pipeline to identifying candidate effectors in *L. maculans*

5.3.1.2 Candidate effector genes

Three candidate effectors were identified upon following the aforementioned pipeline. We were also able to conduct preliminary characterisation of *AvrLm9* gene features.

5.3.1.2.1 Candidate A

This candidate gene was identified as uniquely lost in isolate D8 (Figure 5.3; Chapter 4, Table 4.6) through the gene loss analysis. This gene lies within an AT-rich region containing no other gene annotations, on SC_3 (33.6% GC content of 304 Kb region containing gene; Figure 5.4), and is 228 bp, with no predicted SNPs in its coding region. The annotation associated with this gene was 'predicted protein'.

As per the RNA seq data, this gene was highly upregulated 7 dpi (FPKM values- 7 dpi- 3676.375 and 14 dpi-154.295).

Blast2Go search for this gene resulted in 'predicted protein' being the most informative hit. It also reported this gene shares 41% hit identity to another gene (Candidate B), which is discussed below.

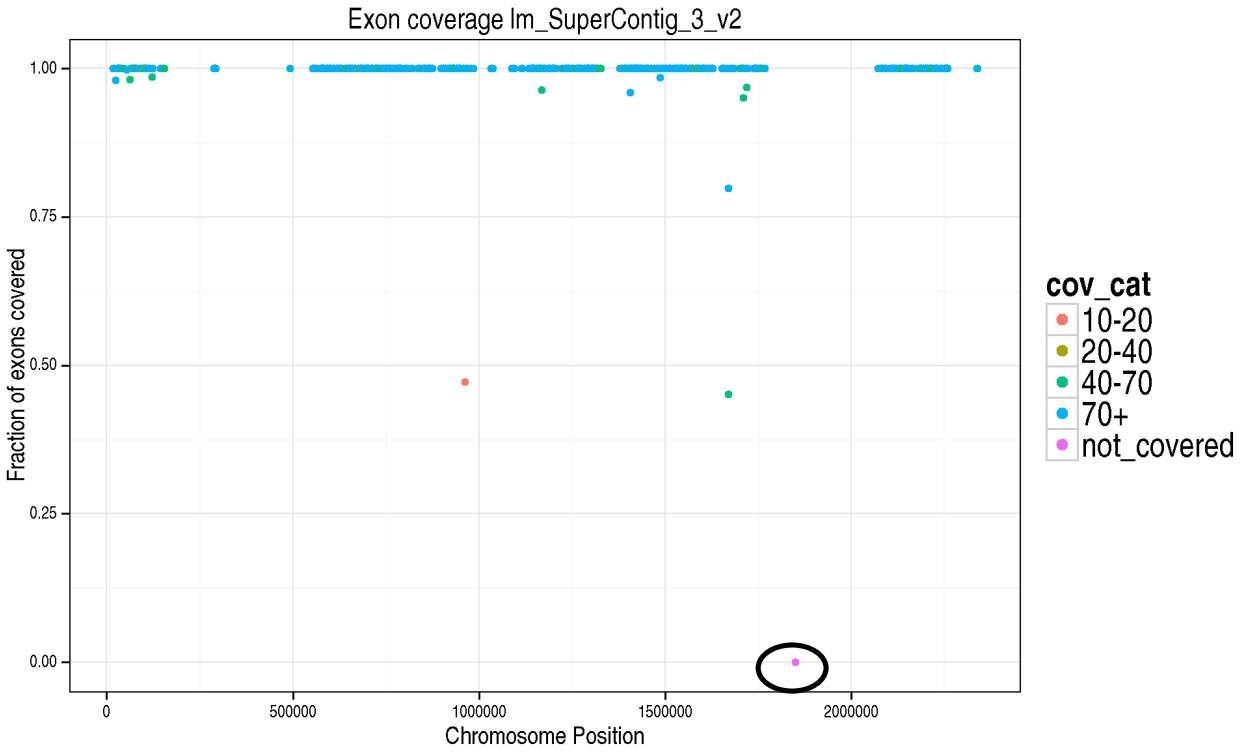

Figure 5.3 Candidate effector gene lost uniquely on SC 3 in isolate D8

(Y axis-Fraction exon of gene is covered by reads; X axis- Position of gene on supercontig; Legend shows coverage categories based on total read coverage of gene; Candidate effector gene circled; The gene was located between positions 1500000 and 2000000)

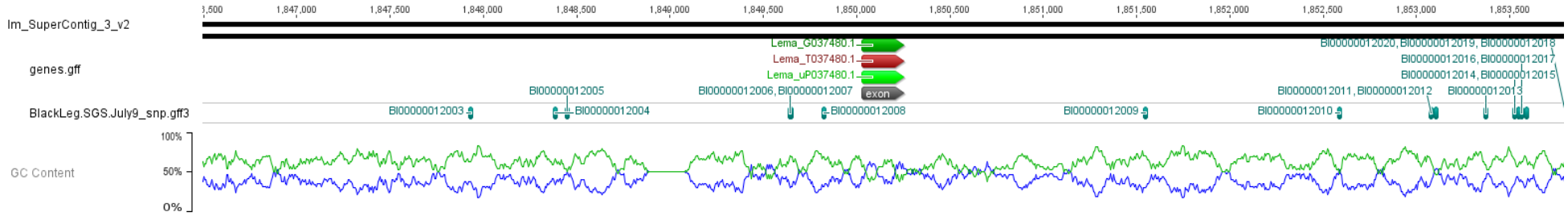

Figure 5.4 Region containing candidate effector gene A on SC_3 in an AT rich intergenic region of 304 Kb containing no other gene predictions (%GC content – Blue line=GC content and Green line=AT content) Image from Geneious (Version 6.1.8; Kearse et al., 2012)

5.3.1.2.2 *Candidate B*

This gene is located on SC_15, in a gene sparse AT-rich region and is 467 bp (Figure 5.5). RNA seq data reports this gene to be highly upregulated 7 dpi (D5 FPKM values- 7 dpi- 2843.285 and 14 dpi- 45.0768). It displays two SNPs within its coding region, one converting a leucine to a proline and the other SNP converting asparagine to histidine. Assuming the default cut-off, PROVEAN predicts both SNPs to be deleterious to protein functionality. This gene does not show any gene loss amongst the isolates analysed and the Blast2Go annotation for it is "predicted protein". D5, the isolate for which RNA seq data is available, does not contain any SNPs in the coding region of this gene (Table 5.1). Interestingly this candidate showed 20.6% identity to *AvrLmJ1* (Van de Wouw et al., 2014).

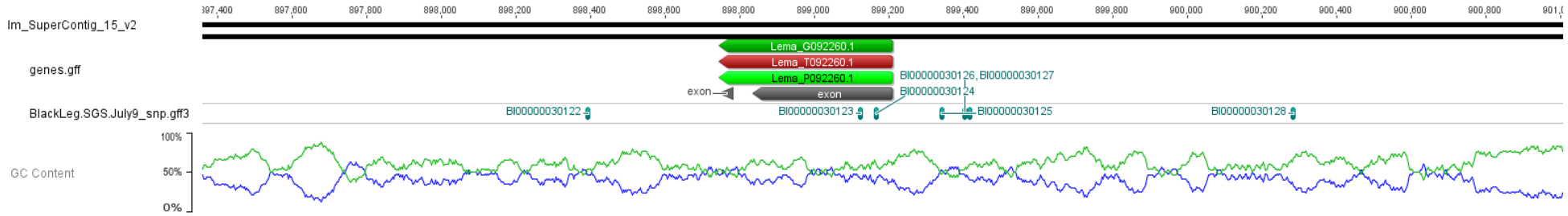

Figure 5.5 Region containing candidate effector gene B on SC_15 in an AT rich intergenic region of 210 Kb containing no other gene predictions (%GC content – Blue line=GC content and Green line=AT content) Image from Geneious (Version 6.1.8; Kearse et al., 2012)

Table 5.1 SNP profile for candidate gene B in our sample set

(SC-SuperContig; Pos-Position on SC; D-Deleterious effect of SNP mutation; PROVEAN score- score of < or = -2.5 is deleterious; Highlighted pink-Isolates containing deleterious SNP)

Pos	SNP	Ref	D1	D2	D3	D4	D5	D6	D8	D9	D12	bp pos	AA pos	AA From	AA To	Effect	PROVEAN score
899126	A/G	A	A	G	A	G	A	A	A	A	A	86	29	L	P	D	-3.5
899169	G/T	T	T	G	T	G	T	T	T	T	T	43	15	N	H	D	-2.5

5.3.1.2.2.1 *Candidate A and B comparison*

A Geneious alignment of translation products of both candidates showed shared homology of the motif seen in Figure 5.6. Overall, both protein products shared 31.2% pairwise identity when aligned using a Geneious alignment (Gap open penalty-12, Gap extension penalty-3). The two products shared 91.7% identity at the probable conserved motif of 'CLCTLXDTLWRM'.

When compared to protein products of all known *AvrLm* genes, *AvrLmJ1* was the only gene that shared some identity at the probable conserved motif between Candidate A and B (Figure 5.7). The overall identity between the three sequences after a Geneious alignment (Gap open penalty 12, gap extension penalty 3) was 20.3%. At the motif site, the identity was 58.3%. Read mapping to the reference for all isolates in this dataset displayed conservation at the motif site for both genes except isolate D8 which loses Candidate A.

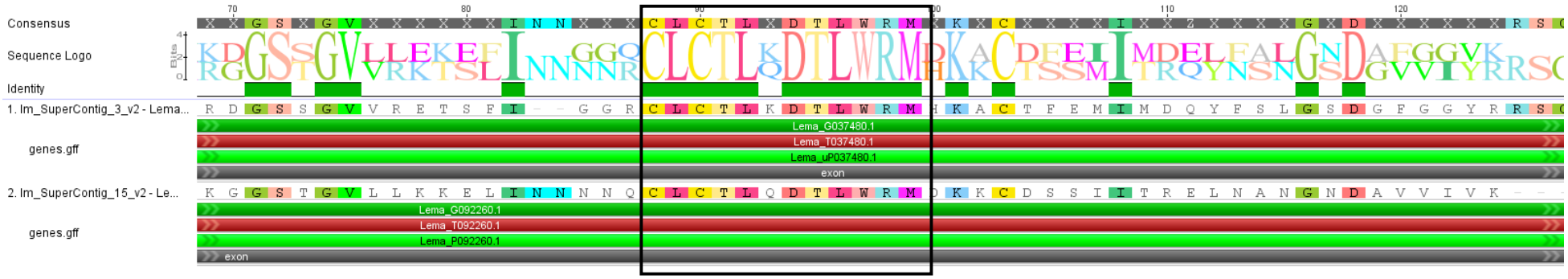

Figure 5.6 Probable conserved motif 'CLCTLXDTLWRM' between Candidate A and Candidate B

(Image from Geneious (Version 6.1.8; Kearse et al., 2012)

(Conserved motif within black square)

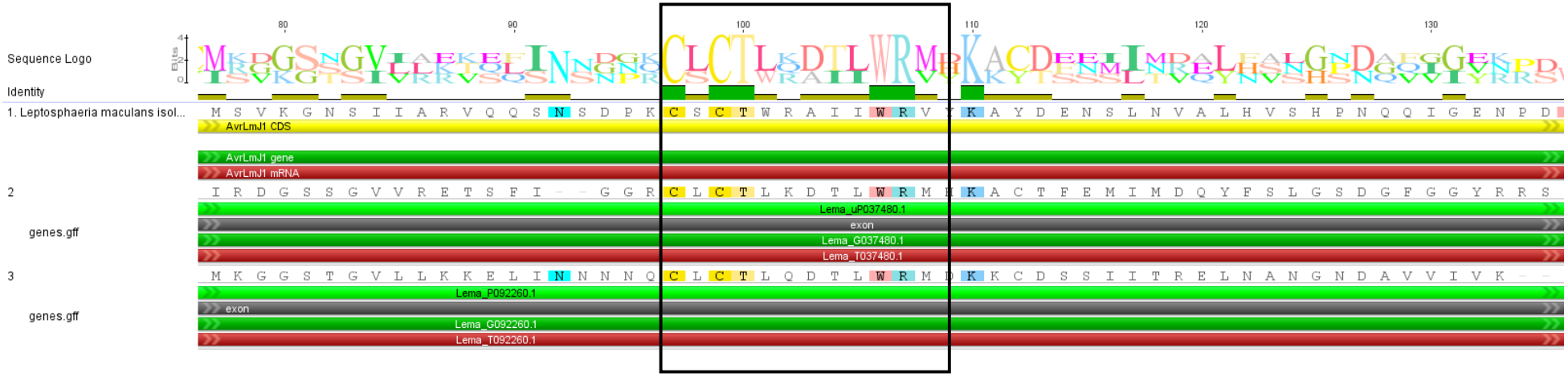

Figure 5.7 Protein product alignment showing identity at probable conserved motif site between *AvrLmJ1* (1), Candidate A (2) and Candidate B (3) (Image from Geneious (Version 6.1.8; Kearse et al., 2012)

(Conserved motif within black square)

5.3.1.2.3 Candidate C for AvrLm5

As per Table 4.3 (Chapter 4, See Section 4.3.3), PAV reads would map uniquely to each isolate in the data set, versus D5, to candidate *AvrLm5* effectors. The only gene that had PAV reads mapping uniquely to isolate D1 and D6 vs D5 was Lema_G004720.1 on SC_0. Incidentally, this gene was also lost uniquely in D5. It is 405 bp, 3.7% of the translated proteins are cysteines and is located in a ~192 kb AT-rich region devoid of other genes but riddled with SNPs (Figure 5.8). This gene does not contain any predicted SNPs in its coding region. Blast2Go analysis of the gene yielded the annotation "predicted protein". We could not correlate the RNA seq expression data because it is lost in D5. If the method to attribute virulence is deletion, this candidate displays appropriate traits to be considered as an *AvrLm5* gene candidate.

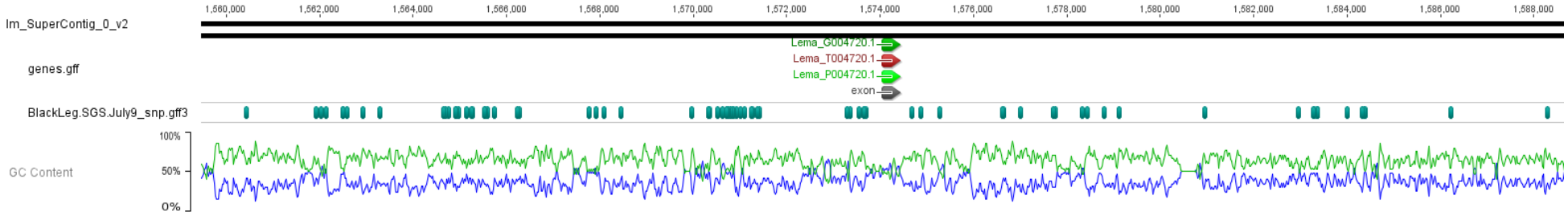

Figure 5.8 Region containing candidate effector gene C on SC_0 in an AT rich intergenic region of 192 Kb containing no other gene predictions (%GC content – Blue line=GC content and Green line=AT content) Image from Geneious (Version 6.1.8; Kearse et al., 2012)

5.3.1.2.4 SC_13 region of interest

SNP predictions (Table 5.2, Zander, 2015) in the region of interest on SC_13 also matched the virulent phenotype for *AvrLm8* in isolates D1, D3, D5 and D9 (See Chapter 4; Table 4.3, Section 4.3.3,). Preliminary analyses also suggest the presence of probable signal peptides in the region. SNP73 (See Chapter 2; SC_13 location 1467717), previously found to be significantly associated with isolating fungal samples from *B. napus* as the stubble species of cultivars in field, was noted to be associated with the absence of this set of thirteen genes in D1, D3, D5 and D9 (Figure 5.9; Table 5.3). It is located in a 1.7 kb region between the genes Lema_G082100.1 and Lema_082110.1. Upon closer inspection, this region was found to contain up to 18 SNPs in each of the aforementioned isolates (18 in D1, 17 in rest), compared to the reference (Table 5.3). Open Reading Frame prediction (ORF) resulted in eleven forward and five reverse frames (Figure 5.9). Initial SignalP (Petersen et al., 2011) analysis run on a six-way translation of the region did not reveal any secreted protein being encoded in the region.

Linkage disequilibrium analysis of SNPs to confirm the association of predicted SNPs in the region containing SNP73 resulted in some blocks of LD between pairs of SNPs (Figure 5.10). SNPs 1,2,4,5,8,9, and 10 (Table 5.2) were noted to all be significantly associated with each other. The same was noted for SNP3 and SNP7. SNP6 was not associated with any other SNP. Initial analyses of these SNPs indicate them to be RIP mutations. Of the SNPs detailed in Table 5.2, SNP73 from previous predictions was not reported by the predictions in Zander (2015) resource due to the difference in the isolates based on which SNP prediction was conducted. Chapter 3 (Zander et al., 2013) used isolates 04MGPS021 and 04MGPP041 for SNP prediction whereas the SNP resource used here (Zander, 2015) used isolates D1, D2, D3, D4, D5, D6, D8, D9, D10 and D12.

Collectively these results indicate that although there may not be a predictable gene in the sequence, there is a possibility that a novel uncharacterised gene that may be present, that is not yet identified in the reference due to misassembly or gaps in the reference genome sequence.

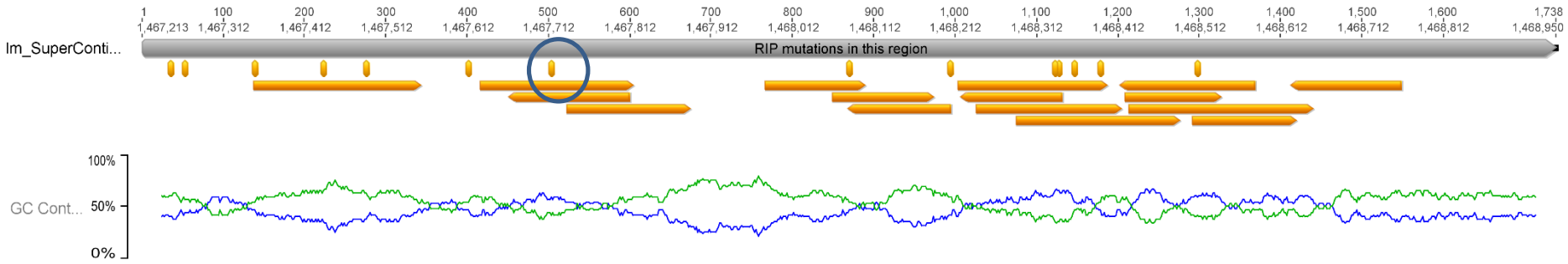

Figure 5.9 Predicted ORFs in the 1.7 Kb region containing SNP 73 on SC_13

(Grey track-SC_13 region of interest; Yellow dots-SNPs in the region (Appendix Table A1); Yellow tracks- Predicted Open Reading Frames (ORF) in the region, forward or reverse direction indicated by arrowhead; SNP73-circled; Image from Geneious (Version 6.1.8; Kearse et al., 2012)

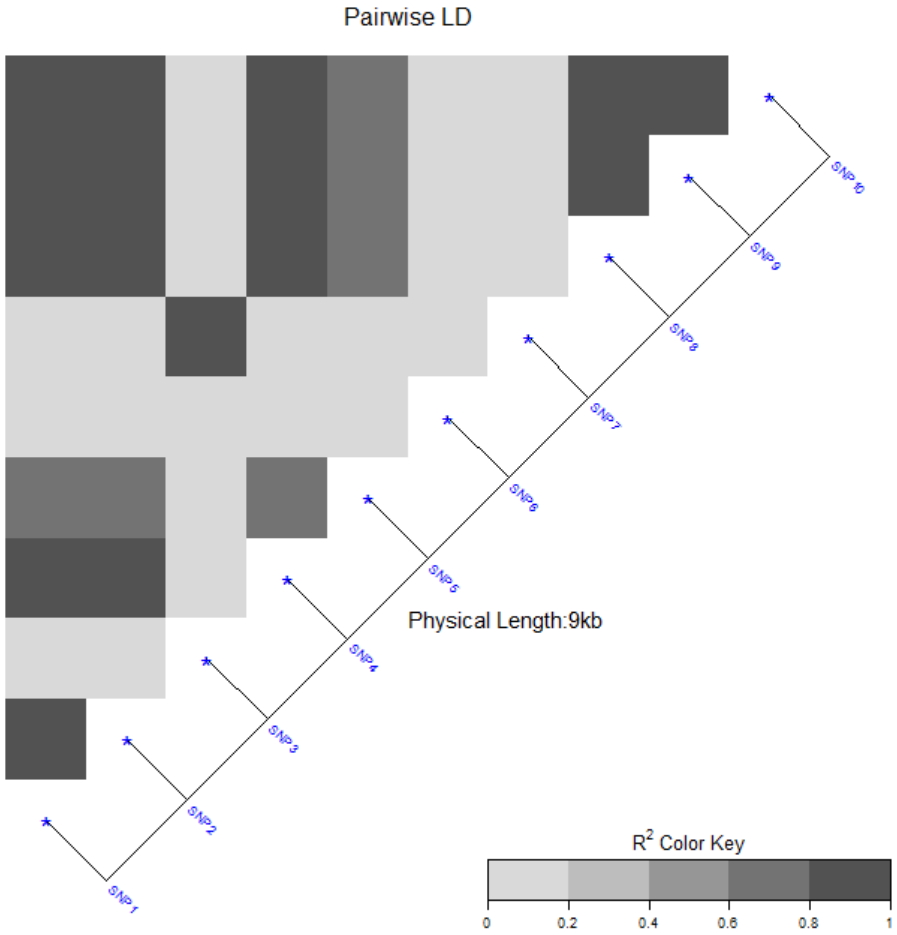

Figure 5.10 Pairwise LD based on R^2 values in SNPs predicted in the region of interest on SC_13 between genes Lema_082100.1 and Lema_082110.1

(0 indicates high recombination; 1 indicates no recombination or linkage disequilibrium)

Table 5.2 Predicted SNPs in region of interest containing potential candidate gene D on SC_13

SNP name	SNP	Position	SNP score	D1	D2	D3	D4	D5	D6	D8	D9	D12
SNP 1	A/G	1,467,352	208	A	G	A	G	A	G	G	A	G
SNP 2	A/G	1,467,436	246	A	G	A	G	A	G	G	A	G
SNP 3	A/G	1,467,552	45	G	G	G	G	G	G	G	G	A
SNP 4	A/G	1,467,614	312	A	G	A	G	A	G	G	A	G
SNP 5	C/T	1,467,930	304	T	C	T	T	T	C	C	T	C
SNP 6	C/G	1,468,011	69	C	C	C	C	C	G	C	C	C
SNP 7	C/T	1,468,013	78	C	C	C	C	C	C	C	C	T
SNP 8	A/G	1,468,391	171	A	G	A	G	A	G	G	A	G
SNP 9	A/G	1,468,666	78	A	G	A	G	A	G	G	A	G
SNP 10	A/G	1,468,685	152	A	G	A	G	A	G	G	A	G

Table 5.3 SNPs in the 1.7 kb region containing SNP73 on SC_13, in isolates D1, D3, D5 and D9

SNP	Position	D1	D3	D5	D9
C > T	1467249	Y	Y	Y	Y
G > A	1467266	Y	Y	Y	Y
G > A	1467352	Y	Y	Y	Y
G > A	1467436	Y	Y	Y	Y
G > A	1467489	Y	Y	Y	Y
G > A	1467614	Y	Y	Y	Y
G > A (SNP73)	1467717	Y	Y	Y	Y
C > G	1468082	Y	N	N	N
G > A	1468207	Y	Y	Y	Y
G > A	1468335	Y	Y	Y	Y
G > A	1468340	Y	Y	Y	Y
G > A	1468359	Y	Y	Y	Y
G > A	1468391	Y	Y	Y	Y
G > A	1468511	Y	Y	Y	Y
G > A	1468631	Y	Y	Y	Y
G > A	1468635	Y	Y	Y	Y
G > A	1468666	Y	Y	Y	Y
G > A	1468685	Y	Y	Y	Y

5.3.1.2.5 AvrLm9

The SNP data was searched for occurrence within candidate effectors in D1. Hypothetically, SNP(s) mutating D1 but not the other isolates or the reference could be diversifying *AvrLm9* candidates in that isolate. Based on prior work, we expect this gene to be located in SC_12 in the AvrLm3-4-7-9 cluster (Rouxel and Balesdent, 2005; Van de Wouw et al., 2010). Therefore, SNPs differing only in D1 as compared to the other isolates including the reference, on SC_12 were explored. This gave us a short list of 30 SNPs down from 1199 SNPs with a SNP score of < or =5. However, only one SNP of these 30 was found to be located within a gene, which was not a suitable candidate for an effector gene. This led us to believe that the *AvrLm9* gene locus undergoes a deletion as the method of diversification as opposed to SNP mutations and thus will not be present in the reference.

5.3.1.3 RNA seq data for candidates

RNA seq data available for isolate D5 (Lowe et al., 2014) was compared to analyse the expression levels for known versus candidate effector genes. Candidate A and B were highly upregulated seven days post inoculation on a susceptible cultivar when compared to the expression of known avirulence genes (Figure 5.11).

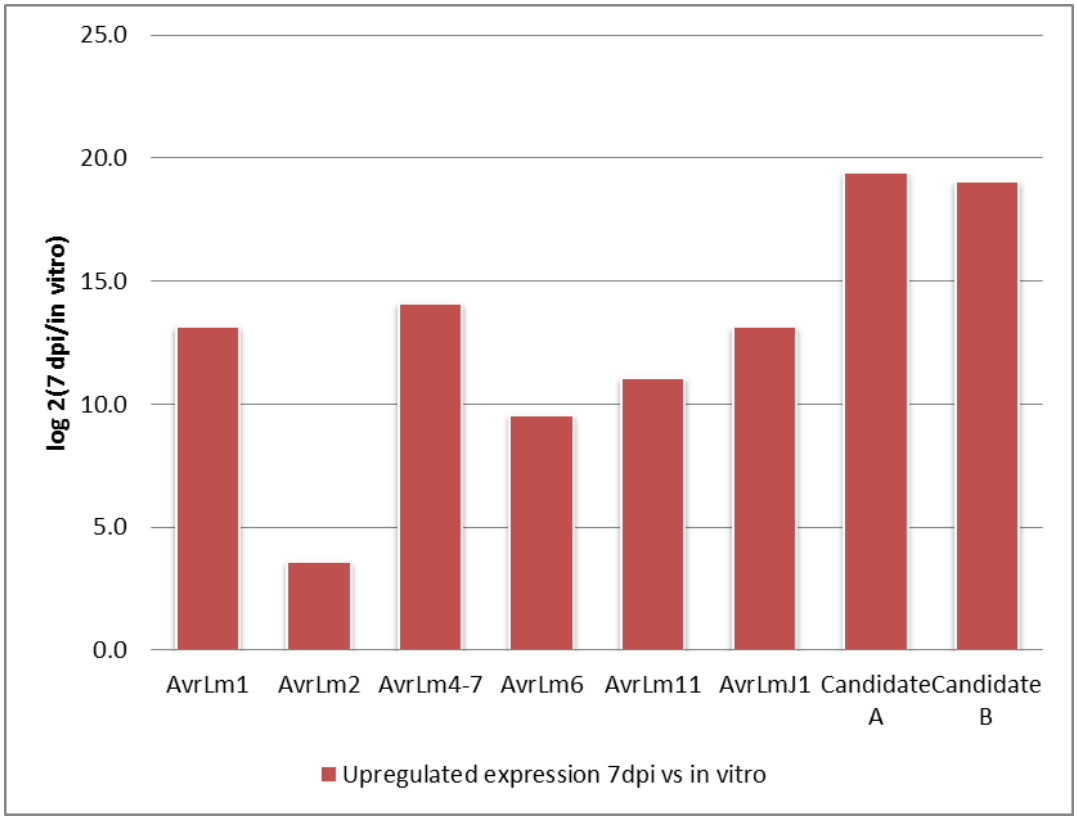

Figure 5.11 Upregulated expression based on RNA seq data of known avirulence gene compared to candidate effector gene A and B

(Vertical axis: Ratio of upregulation of effector genes in plants versus in vitro based on FPKM values)

5.3.2 Genomics using three methods

In order to lay the foundation for building a pan-genome for *L. maculans*, we aimed to combine three different types of genomic data sources. The PAV data resource is vast and on an average, ~2% of total PAV reads map to the reference genome (Zander, 2015). Therefore, we explored the combination of PAV reads, gene loss and SNP data in reference to known annotations in the reference genome as a starting point.

Zander (2015) validated the PAKAP prediction method via PCR amplification of 14 predicted PAV regions in the genome, of which 92.3% predictions were successfully validated. SGSGeneloss was validated by correlating avirulence gene composition of differential isolates (Chapter 4, Golicz et al., 2015). Golicz et al. (2015) additionally validated this method by amplifying the mating-type locus for the samples in their dataset, which was also accurately predicted as lost. Lastly, SGSautoSNP was validated via PCR amplification of 20 randomly chosen SNPs. Of the 20 SNPs, 18 amplified successfully and were found to align 100% with the predictions of the method (Zander et al., 2013). Chapter 3 also validates this method (Patel et al., 2015).

As seen in Figure 5.12, when compared to available genomic annotations in the reference genome, we can further characterise the genomic environment within and around each gene using gene loss and SNP predictions. PAV helped us identify unique genes amongst this data set which was also complemented by gene loss and SNP data. Figure 5.12 shows that we can use PAV data to highlight unique regions in the genome that may otherwise not have any annotated genes but might contain novel genes present in population isolates (Gene A in Figure 5.12).

Collective observation of gene loss, PAV and SNP data led us to understand that the genomic content of each isolate varies significantly to one another (Table 5.4). Furthermore, patterns of SNP mutations and deletion are also unique to each genomic environment. Of the 12,469 predicted genes in the reference genome (Rouxel et al., 2011), 0.13-0.5% of genes displayed PAV gene hits.

D6 and D1 reported the highest number of unique PAV gene hits to known annotations at 63 and 57 genes respectively and D9 and D3 had the least number of gene hits at 17 and 18 genes respectively (Table 5.4). Based on the collation of the number of samples affected by both or either SNPs and gene loss, most samples did not show any significance in the number of genes that were possibly diverged with only gene loss or gene loss and SNPs. As found previously, the majority of genes

highlighted in the data across all isolates were located on SC_13 (Table 5.4, Table 5.5). PAV reads were also observed to map to intergenic locations within the genome.

We were also able to validate the presence of a putative secondary metabolite cluster on SC_13. Table 5.4 genes highlighted purple are thought to belong to the cluster and these were also highlighted by PAVs to be unique in isolates that didn't display gene loss. The majority of gene annotations obtained through Blast2Go (Conesa et al., 2005) were hypothetical or predicted proteins and we were unable to gain more information about their function. Most other gene annotations pertained to housekeeping genes and other processes.

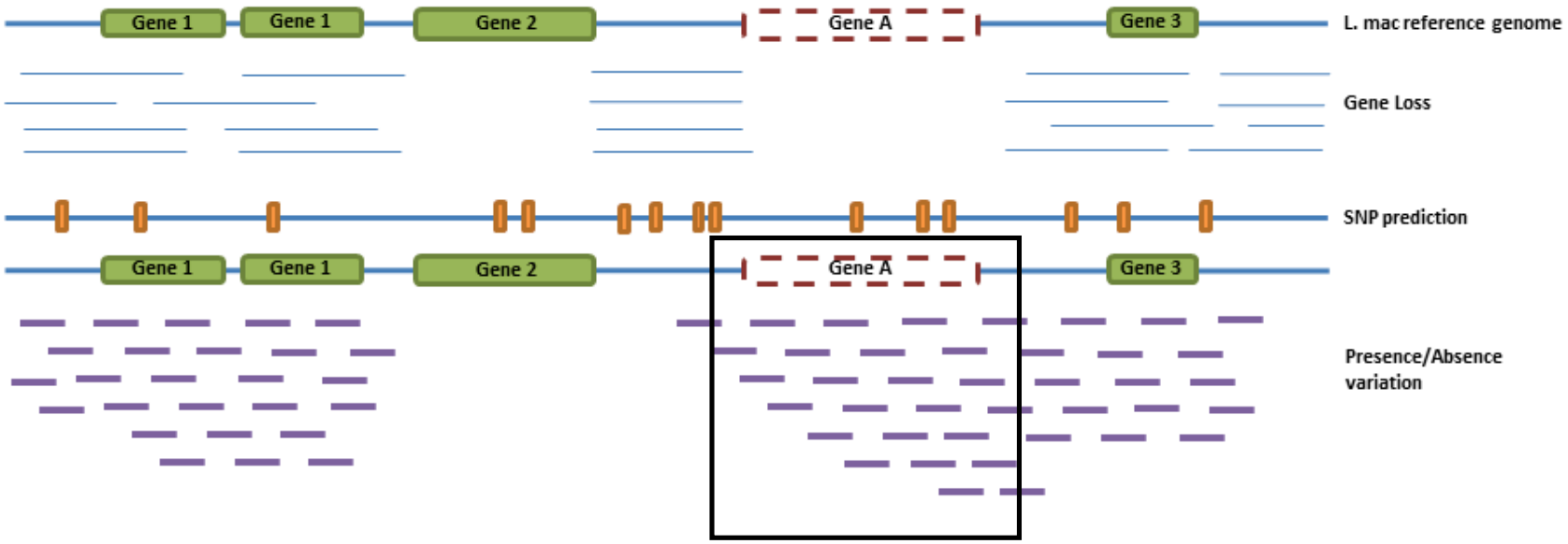

Figure 5.12 Gaps in genomic data for a sample are filled using Gene Loss, PAVs and SNP prediction

(Thin blue lines-Reads mapping to the reference genome from hypothetical isolate X; Gene1, Gene 2 and Gene 3-Annotated genes on the reference genome; Gene A- Gene not predicted on reference genome but present in hypothetical isolate X; Orange rectangles-Predicted SNPs in hypothetical isolate X; Purple dashed lines- PAV reads mapping to the reference genome; Black square- Indicates PAV reads that do not map to the reference but would map to Gene A which is present in Isolate X

Table 5.4 PAV reads unique in isolates D1-D6, D8, D9 and D12 combined with SNP and Gene loss results

(Bl-Blackleg SNPs; Highlighted yellow-Candidate effector gene C; Highlighted purple-Secondary metabolite cluster genes predicted in Chapter 4; Highlighted green-Genes containing SNPs and predicted as lost; Highlighted Pink-Genes do not contain SNPs but are predicted as lost; Highlighted blue-Genes do not contain SNPs and are not predicted as lost)

PAV in	Gene Name	Annotation	SC	Start	End	PAV vs	SNP	Gene Loss
D1	Lema_G004720.1	predicted protein	0	1,574,046	1,574,450	D5	No	D5
D2	Lema_G002200.1	major facilitator superfamily transporter	0	763,884	765,737	D1	Bl789-797	D1, D9
D3	Lema_G002200.1		0	763,884	765,737	D1		
D4	Lema_G002200.1		0	763,884	765,737	D1		
D5	Lema_G002200.1		0	763,884	765,737	D1		
D6	Lema_G002200.1		0	763,884	765,737	D1		
D8	Lema_G002200.1		0	763,884	765,737	D1		
D12	Lema_G002200.1		0	763,884	765,737	D1		
D1	Lema_G013150.1	major facilitator superfamily transporter	1	229,948	233,218	D3, D12	Bl4734-4736	No
D6	Lema_G013150.1		1	229,948	233,218	D12		
D8	Lema_G013150.1		1	229,948	233,218	D12		
D5	Lema_G015080.1	predicted protein	1	838,734	839,305	D6	No	D1, D4, D6, D8, D9
D12	Lema_G015080.1		1	838,734	839,305	D6		
D5	Lema_G015090.1	predicted protein	1	839,417	839,638	D6	No	D1, D4, D6, D8, D9
D12	Lema_G015090.1		1	839,417	839,638	D6		
D2	Lema_G015110.1	predicted protein	1	844,294	846,759	D1	Bl5190-5194	No
D3	Lema_G015110.1		1	844,294	846,759	D1		

D5	Lema_G015110.1		1	844,294	846,759	D1, D6		
D12	Lema_G015110.1		1	844,294	846,759	D1, D6		
D1	Lema_G018140.1	predicted protein	1	1,722,362	1,724,272	D5	Bl5821	No
D1	Lema_G018900.1	predicted protein	1	1,931,202	1,931,996	D3, D4, D9, D12	Bl5927-5929	No
D2	Lema_G018900.1		1	1,931,202	1,931,996	D9		
D6	Lema_G018900.1		1	1,931,202	1,931,996	D4		
D6	Lema_G018900.1		1	1,931,202	1,931,996	D12		
D8	Lema_G018900.1		1	1,931,202	1,931,996	D9		
D8	Lema_G018900.1		1	1,931,202	1,931,996	D12		
D1	Lema_G023430.1	predicted protein	1	3,356,807	3,357,149	D2, D3, D5, D9	No	D3, D5, D8, D9, D12
D1	Lema_G036130.1	predicted protein	3	1,421,067	1,423,417	D5, D12	Bl15505	No
D6	Lema_G036130.1		3	1,421,067	1,423,417	D12		
D8	Lema_G036130.1		3	1,421,067	1,423,417	D12		
D2	Lema_G037120.1	predicted protein	3	1,668,837	1,669,337	D1, D5, D8, D9	Bl11672	No
D4	Lema_G037120.1		3	1,668,837	1,669,337	D1, D5, D6, D8, D9		
D12	Lema_G037120.1		3	1,668,837	1,669,337	D1,D5, D6, D8, D9		
D4	Lema_G037130.1	predicted protein	3	1,669,441	1,669,644	D9		
D12	Lema_G037130.1		3	1,669,441	1,669,644	D9		
D1	Lema_G037220.1	similar to Delta(3,5)-Delta(2,4)-dienoyl-CoA isomerase	3	1,703,798	1,704,842	D6	Bl11707-Bl11714	No
D4	Lema_G037220.1		3	1,703,798	1,704,842	D6		
D5	Lema_G037220.1		3	1,703,798	1,704,842	D6		
D8	Lema_G037220.1		3	1,703,798	1,704,842	D6		

D12	Lema_G037220.1		3	1,703,798	1,704,842	D6		
D1	Lema_G043290.1	dynamin family protein	4	1,428,187	1,430,545	D2, D4, D5, D6, D9, D12	Bl13630	D5
D3	Lema_G043290.1		4	1,428,187	1,430,545	D2		
D6	Lema_G043290.1		4	1,428,187	1,430,545	D2		
D8	Lema_G043290.1		4	1,428,187	1,430,545	D2, D6, D12		
D9	Lema_G043290.1		4	1,428,187	1,430,545	D2		
D12	Lema_G043290.1		4	1,428,187	1,430,545	D2		
D1	Lema_G043360.1	predicted protein	4	1,445,551	1,447,257	D4	Bl13640, Bl13642	No
D1	Lema_G050760.1	predicted protein	5	347,544	347,994	D12	No	D12
D2	Lema_G050760.1		5	347,544	347,994	D12		
D3	Lema_G050760.1		5	347,544	347,994	D12		
D4	Lema_G050760.1	predicted protein	5	347,544	347,994	D12		
D5	Lema_G050760.1		5	347,544	347,994	D12		
D6	Lema_G050760.1		5	347,544	347,994	D12		
D9	Lema_G050760.1		5	347,544	347,994	D12		
D1	Lema_G050770.1	predicted protein	5	348,537	356,093	D12	No	No
D6	Lema_G050770.1		5	348,537	356,093	D12		
D8	Lema_G050770.1		5	348,537	356,093	D12		
D1	Lema_G051020.1	predicted protein	5	415,359	416,426	D4	Bl16986-16987	No
D6	Lema_G051020.1		5	415,359	416,426	D4		
D1	Lema_G051290.1	predicted protein	5	490,652	491,090	D6	No	No
D4	Lema_G051290.1		5	490,652	491,090	D6		
D5	Lema_G051290.1		5	490,652	491,090	D6		

D8	Lema_G051290.1		5	490,652	491,090	D6		
D12	Lema_G051290.1		5	490,652	491,090	D6		
D2	Lema_G053310.1	very large low complexity protein	5	1,055,685	1,065,356	D1, D3, D4, D9	Bl17493-17494	D1, D3, D4, D9
D5	Lema_G053310.1		5	1,055,685	1,065,356	D1, D3, D4, D9		
D6	Lema_G053310.1		5	1,055,685	1,065,356	D1,D3,D4, D9		
D8	Lema_G053310.1		5	1,055,685	1,065,356	D1, D3, D4, D9		
D12	Lema_G053310.1		5	1,055,685	1,065,356	D1, D3, D4, D9		
D1	Lema_G046870.1	predicted protein	6	590,695	591,212	D8, D12	Bl14484	No
D5	Lema_G049660.1	*AvrLm1*	6	1,607,018	1,607,681	D1, D2, D4, D9, D12	Bl15873	D1, D2, D3, D4, D8, D9, D12
D6	Lema_G049660.1		6	1,607,018	1,607,681	D1,D2,D3, D4,D8,D9, D12		
D8	Lema_G049660.1		6	1,607,018	1,607,681	D12		
D1	Lema_G049920.1	predicted protein	6	1,837,314	1,837,498	D3, D5	No	D3, D5
D1	Lema_G049930.1	predicted protein	6	1,837,589	1,837,925	D3, D5	No	D3, D5
D1	Lema_G049940.1	*AvrLm6*	6	1,839,097	1,839,677	D3	No	D3, D5
D6	Lema_G049940.1		6	1,839,097	1,839,677	D3		
D8	Lema_G049940.1		6	1,839,097	1,839,677	D3		
D12	Lema_G049940.1		6	1,839,097	1,839,677	D3		
D1	Lema_G067720.1	predicted protein	7	142,280	143,248	D4, D6, D9, D12	Bl21490	No
D2	Lema_G067720.1		7	142,280	143,248	D9		
D3	Lema_G067720.1		7	142,280	143,248	D9		

D5	Lema_G067720.1		7	142,280	143,248	D6, D9		
D8	Lema_G067720.1		7	142,280	143,248	D6, D9		
D1	Lema_G068460.1	beta-glucosidase	7	331,214	333,121	D5	Bl21566	D5
D1	Lema_G061030.1		8	1,427,511	1,429,907	D9	No	D9
D1	Lema_G076680.1	predicted protein	10	1,575,621	1,576,371	D8, D9	Bl24470	No
D6	Lema_G076680.1		10	1,575,621	1,576,371	D8		
D1	Lema_G082370.1	predicted protein	12	109,077	109,426	D3, D6, D8, D12	No	D3, D6, D8, D12
D4	Lema_G082370.1		12	109,077	109,426	D6		
D5	Lema_G082370.1		12	109,077	109,426	D6		
D1	Lema_G082380.1	predicted protein	12	110,629	113,389	D3, D6, D8, D12	No	D3, D6, D8, D12
D2	Lema_G082380.1		12	110,629	113,389	D3, D6, D12		
D4	Lema_G082380.1		12	110,629	113,389	D3, D6, D12		
D5	Lema_G082380.1		12	110,629	113,389	D3, D6, D12		
D9	Lema_G082380.1		12	110,629	113,389	D3, D6, D12		
D1	Lema_G086300.1	predicted protein	12	1,400,949	1,401,849	D3	No	D3
D1	Lema_G086310.1	predicted protein	12	1,402,878	1,403,303	D3	No	D3
D1	Lema_G078330.1		13	319,292	320,025	D2	Bl24876	No
D6	Lema_G081630.1	predicted protein	13	1,350,995	1,351,795	D12	Bl25690	D1, D3, D5, D9
D8	Lema_G081630.1		13	1,350,995	1,351,795	D12		
D2	Lema_G081640.1	predicted protein	13	1,352,946	1,353,440	D1, D3, D12	No	D1, D3, D4, D5, D8, D9, D12
D6	Lema_G081640.1		13	1,352,946	1,353,440	D1,D3,D4, D12		

D8	Lema_G081640.1		13	1,352,946	1,353,440	D12		
D6	Lema_G081650.1	predicted protein	13	1,353,777	1,356,233	D4,D8,D12	Bl25692	No
D8	Lema_G081650.1		13	1,353,777	1,356,233	D12		
D1	Lema_G081880.1	*SirH*	13	1,411,646	1,412,969	D4	Bl25682	D4
D2	Lema_G081880.1		13	1,411,646	1,412,969	D4		
D3	Lema_G081880.1		13	1,411,646	1,412,969	D4		
D5	Lema_G081880.1		13	1,411,646	1,412,969	D4		
D6	Lema_G081880.1		13	1,411,646	1,412,969	D4		
D8	Lema_G081880.1		13	1,411,646	1,412,969	D4		
D9	Lema_G081880.1		13	1,411,646	1,412,969	D4		
D12	Lema_G081880.1		13	1,411,646	1,412,969	D4		
D1	Lema_G081890.1	Predicted protein	13	1,413,647	1,413,827	D4	No	D4
D6	Lema_G081890.1		13	1,413,647	1,413,827	D4		
D1	Lema_G081900.1	*LmUVI-1h*	13	1,415,843	1,416,552	D4	Bl25683	D4
D6	Lema_G081900.1		13	1,415,843	1,416,552	D4		
D1	Lema_G081910.1	Predicted protein	13	1,417,543	1,418,021	D4	Bl25684	D4
D6	Lema_G081910.1		13	1,417,543	1,418,021	D4		
D1	Lema_G081920.1	Polyketide synthase	13	1,418,093	1,424,481	D4, D8	Bl25685-25689	D4
D2	Lema_G081920.1		13	1,418,093	1,424,481	D4		
D3	Lema_G081920.1		13	1,418,093	1,424,481	D4		
D5	Lema_G081920.1		13	1,418,093	1,424,481	D4		
D6	Lema_G081920.1		13	1,418,093	1,424,481	D4		
D6	Lema_G081920.1		13	1,418,093	1,424,481	D8		
D8	Lema_G081920.1		13	1,418,093	1,424,481	D4		
D9	Lema_G081920.1		13	1,418,093	1,424,481	D4		
D12	Lema_G081920.1		13	1,418,093	1,424,481	D4		

D1	Lema_G081930.1	*LmHDX1*	13	1,424,869	1,426,310	D8	Bl25690	D8
D2	Lema_G081930.1		13	1,424,869	1,426,310	D8		
D3	Lema_G081930.1		13	1,424,869	1,426,310	D8		
D4	Lema_G081930.1		13	1,424,869	1,426,310	D8		
D5	Lema_G081930.1		13	1,424,869	1,426,310	D8		
D6	Lema_G081930.1		13	1,424,869	1,426,310	D8		
D9	Lema_G081930.1		13	1,424,869	1,426,310	D8		
D12	Lema_G081930.1		13	1,424,869	1,426,310	D8		
D2	Lema_G081940.1	Aspartate aminotransferase	13	1,426,780	1,428,165	D1, D8	No	D1, D3, D5, D8, D9
D4	Lema_G081940.1		13	1,426,780	1,428,165	D1, D5, D8		
D6	Lema_G081940.1		13	1,426,780	1,428,165	D1,D8		
D12	Lema_G081940.1		13	1,426,780	1,428,165	D1, D8		
D2	Lema_G081950.1	Aflatoxin biosynthesis ketoreductase	13	1,428,596	1,429,536	D1, D3, D9	Bl25692	D1, D3, D5, D8, D9
D4	Lema_G081950.1		13	1,428,596	1,429,536	D1, D3, D9		
D6	Lema_G081950.1		13	1,428,596	1,429,536	D1,D3,D8, D9		
D12	Lema_G081950.1		13	1,428,596	1,429,536	D1, D3, D9		
D2	Lema_G081960.1	Stress responsive a b barrel domain protein	13	1,430,732	1,431,163	D1, D5	Bl25700-25704	D1, D3, D5, D8, D9
D4	Lema_G081960.1		13	1,430,732	1,431,163	D1, D5		
D6	Lema_G081960.1		13	1,430,732	1,431,163	D1		
D6	Lema_G081960.1		13	1,430,732	1,431,163	D8		
D2	Lema_G081980.1	Salicylate hydroxylase	13	1,438,204	1,439,836	D9	Bl25724-Bl25738	D1, D3, D5, D9
D4	Lema_G081980.1		13	1,438,204	1,439,836	D9		
D6	Lema_G081980.1		13	1,438,204	1,439,836	D9		
D8	Lema_G081980.1		13	1,438,204	1,439,836	D9		
D2	Lema_G081990.1	Predicted protein	13	1,440,856	1,441,392	D1, D5	No	D1, D3,

D4	Lema_G081990.1		13	1,440,856	1,441,392	D1, D5		D5, D9
D6	Lema_G081990.1		13	1,440,856	1,441,392	D1		
D8	Lema_G081990.1		13	1,440,856	1,441,392	D1		
D8	Lema_G081990.1		13	1,440,856	1,441,392	D5		
D12	Lema_G081990.1		13	1,440,856	1,441,392	D1, D5		
D2	Lema_G082000.1	Short-chain dehydrogenase reductase sdr	13	1,441,845	1,442,615	D1, D3, D5	No	D1, D3, D5, D9
D4	Lema_G082000.1		13	1,441,845	1,442,615	D1, D3, D5		
D6	Lema_G082000.1		13	1,441,845	1,442,615	D1		
D8	Lema_G082000.1		13	1,441,845	1,442,615	D1, D3, D5		
D12	Lema_G082000.1		13	1,441,845	1,442,615	D1, D3, D5		
D2	Lema_G082010.1	Zinc-binding oxidoreductase	13	1,444,412	1,445,637	D1	Bl25753-25757	D1, D3, D5, D9
D4	Lema_G082010.1		13	1,444,412	1,445,637	D1		
D6	Lema_G082010.1		13	1,444,412	1,445,637	D1		
D8	Lema_G082010.1		13	1,444,412	1,445,637	D1		
D12	Lema_G082010.1		13	1,444,412	1,445,637	D1		
D2	Lema_G082020.1	O-methyltransferase	13	1,446,238	1,447,610	D1, D9	Bl25758-25762	D1, D3, D5, D9
D4	Lema_G082020.1		13	1,446,238	1,447,610	D1, D5, D9		
D6	Lema_G082020.1		13	1,446,238	1,447,610	D1,D9		
D8	Lema_G082020.1		13	1,446,238	1,447,610	D1, D5, D9		
D12	Lema_G082020.1		13	1,446,238	1,447,610	D1, D5, D9		
D2	Lema_G082030.1	Trihydroxytoluene oxygenase	13	1,448,479	1,449,713	D1, D3, D5, D9	Bl25764	D1, D3, D5, D9
D4	Lema_G082030.1		13	1,448,479	1,449,713	D1, D3, D5, D9		
D6	Lema_G082030.1		13	1,448,479	1,449,713	D1,D3,D9		
D8	Lema_G082030.1		13	1,448,479	1,449,713	D1, D3, D5, D9		

D12	Lema_G082030.1		13	1,448,479	1,449,713	D1, D3, D5, D9		
D2	Lema_G082040.1	Predicted protein	13	1,449,952	1,450,340	D3	Bl25767,25768	D1, D3, D5, D9
D4	Lema_G082040.1		13	1,449,952	1,450,340	D3		
D6	Lema_G082040.1		13	1,449,952	1,450,340	D3		
D8	Lema_G082040.1		13	1,449,952	1,450,340	D3		
D12	Lema_G082040.1		13	1,449,952	1,450,340	D3		
D2	Lema_G082050.1	Fatty acid synthase subunit alpha	13	1,451,035	1,454,379	D1, D3, D5, D9	Bl25772-25789	D1, D3, D5, D9
D4	Lema_G082050.1		13	1,451,035	1,454,379	D1, D3, D5, D9		
D6	Lema_G082050.1		13	1,451,035	1,454,379	D1,D3		
D8	Lema_G082050.1		13	1,451,035	1,454,379	D1, D3, D5, D9		
D12	Lema_G082050.1		13	1,451,035	1,454,379	D1, D3, D5		
D2	Lema_G082090.1	Similar to fatty acid synthase beta subunit	13	1,459,766	1,465,599	D9	Bl25825-25837	No
D4	Lema_G082090.1		13	1,459,766	1,465,599	D5, D9		
D6	Lema_G082090.1		13	1,459,766	1,465,599	D9		
D12	Lema_G082090.1		13	1,459,766	1,465,599	D5, D9		
D2	Lema_G082100.1	Fatty acid synthase subunit alpha	13	1,465,862	1,467,212	D1, D5	Bl25838-25841	D1, D3, D5, D9
D4	Lema_G082100.1		13	1,465,862	1,467,212	D1, D5, D9		
D6	Lema_G082100.1		13	1,465,862	1,467,212	D1,D9		
D8	Lema_G082100.1		13	1,465,862	1,467,212	D1, D5		
D12	Lema_G082100.1		13	1,465,862	1,467,212	D1, D5		
D1	Lema_G093960.1	predicted protein	14	85,642	86,377	D12	No	No
D6	Lema_G093960.1		14	85,642	86,377	D12		
D8	Lema_G093960.1		14	85,642	86,377	D12		

D2	Lema_G094000.1	dynamin family protein	14	152,839	155,417	D1	No	D1
D3	Lema_G094000.1		14	152,839	155,417	D1		
D4	Lema_G094000.1		14	152,839	155,417	D1		
D5	Lema_G094000.1		14	152,839	155,417	D1		
D6	Lema_G094000.1		14	152,839	155,417	D1		
D8	Lema_G094000.1		14	152,839	155,417	D1		
D9	Lema_G094000.1		14	152,839	155,417	D1		
D12	Lema_G094000.1		14	152,839	155,417	D1		
D2	Lema_G094010.1	predicted protein	14	157,792	158,001	D1	No	D1
D3	Lema_G094010.1		14	157,792	158,001	D1		
D4	Lema_G094010.1		14	157,792	158,001	D1		
D5	Lema_G094010.1		14	157,792	158,001	D1		
D6	Lema_G094010.1		14	157,792	158,001	D1		
D8	Lema_G094010.1		14	157,792	158,001	D1		
D9	Lema_G094010.1		14	157,792	158,001	D1		
D12	Lema_G094010.1		14	157,792	158,001	D1		
D2	Lema_G094020.1	p-loop containing nucleoside triphosphate hydrolase	14	158,218	167,699	D1	Bl31193-31197	D1
D3	Lema_G094020.1		14	158,218	167,699	D1		
D4	Lema_G094020.1		14	158,218	167,699	D1		
D5	Lema_G094020.1		14	158,218	167,699	D1		
D6	Lema_G094020.1		14	158,218	167,699	D1		
D8	Lema_G094020.1		14	158,218	167,699	D1		
D9	Lema_G094020.1		14	158,218	167,699	D1		
D12	Lema_G094020.1		14	158,218	167,699	D1		
D2	Lema_G098080.1	serine threonine protein kinase	14	1,291,448	1,292,362	D1, D4	No	D1, D3, D8
D5	Lema_G098080.1		14	1,291,448	1,292,362	D1, D4		

D6	Lema_G098080.1		14	1,291,448	1,292,362	D1,D4,D8		
D9	Lema_G098080.1		14	1,291,448	1,292,362	D1, D4		
D12	Lema_G098080.1		14	1,291,448	1,292,362	D1, D4		
D2	Lema_G098090.1	predicted protein	14	1,292,449	1,292,811	D1	No	D1, D3, D8
D5	Lema_G098090.1		14	1,292,449	1,292,811	D1		
D6	Lema_G098090.1		14	1,292,449	1,292,811	D1,D4,D8		
D9	Lema_G098090.1		14	1,292,449	1,292,811	D1		
D12	Lema_G098090.1		14	1,292,449	1,292,811	D1		
D1	Lema_G090930.1	predicted protein	15	500,031	500,476	D6, D9, D12	Bl29804	No
D8	Lema_G090930.1		15	500,031	500,476	D6		
D2	Lema_G091180.1	vegetative incompatibility protein het-e-1	15	555,197	558,578	D1	No	D1
D3	Lema_G091180.1		15	555,197	558,578	D1		
D4	Lema_G091180.1		15	555,197	558,578	D1		
D5	Lema_G091180.1		15	555,197	558,578	D1		
D6	Lema_G091180.1		15	555,197	558,578	D1		
D8	Lema_G091180.1		15	555,197	558,578	D1		
D9	Lema_G091180.1		15	555,197	558,578	D1		
D12	Lema_G091180.1		15	555,197	558,578	D1		
D3	Lema_G091280.1	similar to ribonuclease H	15	585,830	586,321	D1, D2	Bl29847-Bl29848	No
D6	Lema_G091280.1		15	585,830	586,321	D1,D2,D12		
D8	Lema_G091280.1		15	585,830	586,321	D1, D2, D12		
D6	Lema_G104250.1	predicted protein	16	511,036	512,849	D8	Bl34120-34123	No
D1	Lema_G104260.1	het domain protein	16	513,425	514,524	D2, D5, D8	No	D2, D5, D8
D3	Lema_G104260.1		16	513,425	514,524	D2, D5, D8		

D4	Lema_G104260.1		16	513,425	514,524	D2, D5, D8		
D6	Lema_G104260.1		16	513,425	514,524	D2,D8		
D9	Lema_G104260.1		16	513,425	514,524	D2, D5, D8		
D12	Lema_G104260.1		16	513,425	514,524	D2, D5, D8		
D1	Lema_G104270.1	predicted protein	16	514,658	516,407	D2, D5	Bl34125-34130	No
D3	Lema_G104270.1		16	514,658	516,407	D2		
D4	Lema_G104270.1		16	514,658	516,407	D2		
D6	Lema_G104270.1		16	514,658	516,407	D2		
D9	Lema_G104270.1		16	514,658	516,407	D2		
D12	Lema_G104270.1		16	514,658	516,407	D2		
D1	Lema_G104290.1	predicted protein	16	519,580	522,064	D2, D5, D8	Bl34142-34144	No
D6	Lema_G104290.1		16	519,580	522,064	D8		
D1	Lema_G106570.1	predicted protein	16	1,341,710	1,342,574	D4, D6, D9	Bl34977,78,80	No
D5	Lema_G106570.1		16	1,341,710	1,342,574	D6		
D8	Lema_G106570.1		16	1,341,710	1,342,574	D6		
D12	Lema_G106570.1		16	1,341,710	1,342,574	D6		
D1	Lema_G106610.1	tetracycline-efflux transporter	18	31,059	32,963	D4	Bl35280	D4
D1	Lema_G107350.1	predicted protein	18	347,790	348,175	D4	Bl35547,48	No
D2	Lema_G109260.1	heterokaryon incompatibility protein	18	905,419	907,083	D1	Bl36066-Bl36070	D1, D9
D3	Lema_G109260.1		18	905,419	907,083	D1		
D4	Lema_G109260.1		18	905,419	907,083	D1		
D5	Lema_G109260.1		18	905,419	907,083	D1		
D4	Lema_G109270.1	predicted protein	18	909,482	910,569	D6	Bl26071	No
D5	Lema_G109270.1		18	909,482	910,569	D6		
D6	Lema_G106610.1	similar to tetracycline-	18	31,059	32,963	D4	Bl35280	D4

		efflux transporter						
D1	Lema_G111470.1	predicted protein	19	516,753	518,986	D12	No	No
D6	Lema_G111470.1		19	516,753	518,986	D12		
D8	Lema_G111470.1		19	516,753	518,986	D12		
D2	Lema_G113050.1	tpr domain protein	19	956,634	958,686	D1, D12	No	D1, D3, D4, D5, D6, D8, D9, D12
D2	Lema_G113060.1	predicted protein	19	959,197	960,108	D1, D3, D5, D12	No	D1, D3, D4, D5, D6, D8, D9, D12
D1	Lema_G113150.1	predicted protein	20	57,305	58,784	D12	No	D12
D2	Lema_G113150.1		20	57,305	58,784	D12		
D3	Lema_G113150.1		20	57,305	58,784	D12		
D4	Lema_G113150.1		20	57,305	58,784	D12		
D5	Lema_G113150.1		20	57,305	58,784	D12		
D6	Lema_G113150.1		20	57,305	58,784	D12		
D8	Lema_G113150.1		20	57,305	58,784	D12		
D9	Lema_G113150.1		20	57,305	58,784	D12		
D1	Lema_G113160.1	fructosamine-3-kinase	20	69,195	70,206	D12	Bl38836	D12
D6	Lema_G113160.1		20	69,195	70,206	D12		
D8	Lema_G113160.1		20	69,195	70,206	D12		
D1	Lema_G113170.1	cutinase	20	70,380	71,347	D12	No	D12
D6	Lema_G113170.1		20	70,380	71,347	D12		
D8	Lema_G113170.1		20	70,380	71,347	D12		

D1	Lema_G113180.1	fad binding domain protein	20	72,098	73,500	D12	No	D12
D6	Lema_G113180.1		20	72,098	73,500	D12		
D8	Lema_G113180.1		20	72,098	73,500	D12		
D1	Lema_G113190.1	predicted protein	20	201,217	201,836	D6	No	D6
D4	Lema_G113190.1		20	201,217	201,836	D6		
D5	Lema_G113190.1		20	201,217	201,836	D6		
D8	Lema_G113190.1		20	201,217	201,836	D6		
D12	Lema_G113190.1		20	201,217	201,836	D6		
D1	Lema_G114270.1	predicted protein	20	529,112	533,279	D3, D9	Bl39740-42	No
D5	Lema_G114270.1		20	529,112	533,279	D3		
D6	Lema_G114270.1		20	529,112	533,279	D3		
D8	Lema_G114270.1		20	529,112	533,279	D3		
D12	Lema_G114270.1		20	529,112	533,279	D3		
D1	Lema_G114280.1	predicted protein	20	534,387	534,664	D3, D4, D9	No	D3, D4, D9
D2	Lema_G114280.1		20	534,387	534,664	D9		
D5	Lema_G114280.1		20	534,387	534,664	D9		
D6	Lema_G114280.1		20	534,387	534,664	D4,D9		
D8	Lema_G114280.1		20	534,387	534,664	D9		
D12	Lema_G114280.1		20	534,387	534,664	D9		
D1	Lema_G114290.1	predicted protein	20	535,205	538,183	D3, D4, D9	Bl39743-45, 47	No
D2	Lema_G114290.1		20	535,205	538,183	D4, D9		
D5	Lema_G114290.1		20	535,205	538,183	D4, D9		
D6	Lema_G114290.1		20	535,205	538,183	D4,D9		
D8	Lema_G114290.1		20	535,205	538,183	D4, D9		
D12	Lema_G114290.1		20	535,205	538,183	D4, D9		
D1	Lema_G114370.1	similar to Rho GTPase	20	549,910	552,133	D3, D4, D9	Bl39773-	No

D6	Lema_G114370.1	activator (Rgd1)	20	549,910	552,133	D4	39783	
D1	Lema_G114380.1	pyridoxamine phosphate oxidase	20	552,415	553,355	D3, D4, D9	No	D3, D4, D9
D2	Lema_G114380.1		20	552,415	553,355	D9		
D5	Lema_G114380.1		20	552,415	553,355	D9		
D6	Lema_G114380.1		20	552,415	553,355	D4,D9		
D8	Lema_G114380.1		20	552,415	553,355	D9		
D12	Lema_G114380.1		20	552,415	553,355	D9		
D1	Lema_G114390.1	mating type 1-2 protein	20	554,610	555,882	D3, D4, D9	No	D3, D4, D9
D6	Lema_G114390.1		20	554,610	555,882	D4		
D1	Lema_G116710.1	glycosyltransferase family 34 protein	21	263,332	263,889	D4	Bl40702	D4
D2	Lema_G116710.1		21	263,332	263,889	D4		
D3	Lema_G116710.1		21	263,332	263,889	D4		
D5	Lema_G116710.1		21	263,332	263,889	D4		
D6	Lema_G116710.1		21	263,332	263,889	D4		
D8	Lema_G116710.1		21	263,332	263,889	D4		
D9	Lema_G116710.1		21	263,332	263,889	D4		
D12	Lema_G116710.1		21	263,332	263,889	D4		
D1	Lema_G119050.1	predicted protein	22	105,263	106,588	D3	Bl41424-27	No
D1	Lema_G119640.1	predicted protein	23	112,460	115,068	D2, D3, D6, D12	No	D2, D3, D6, D12
D4	Lema_G119640.1		23	112,460	115,068	D2, D3, D6, D12		
D5	Lema_G119640.1		23	112,460	115,068	D2, D3, D6, D12		
D8	Lema_G119640.1		23	112,460	115,068	D2, D3, D6		
D9	Lema_G119640.1		23	112,460	115,068	D2, D3, D12		

D1	Lema_G119650.1	predicted protein	23	115,658	115,852	D2, D12	No	D2, D3, D6, D12
D1	Lema_G119660.1	predicted protein	23	116,196	117,176	D2, D3, D6, D12	No	No
D4	Lema_G119660.1		23	116,196	117,176	D6		
D5	Lema_G119660.1		23	116,196	117,176	D6		
D8	Lema_G119660.1		23	116,196	117,176	D6		

Table 5.5 Statistics of data compilation in this data set

(Gene loss and SNPs-Gene displaying loss and have SNPs in its coding region, where present; Only Gene loss-Genes displaying only loss and contain no predicted SNPs; Only SNPs-Genes containing only SNPs in their coding region and are not reported as lost in this data set; None- Genes neither lost or containing SNPs; Max SC hits-The supercontig containing the most number of genes with PAV hits in each isolate)

Isolate	Total PAV gene hits	Gene loss and SNPs	Only Gene loss	Only SNPs	None	Max SC hits
D1	57	10	22	21	4	20
D2	39	17	16	6	0	13
D3	18	8	6	4	0	13,14
D4	40	15	15	9	1	13
D5	34	9	14	9	2	20
D6	63	26	20	14	3	13
D8	49	17	16	11	5	13
D9	17	6	10	1	0	14
D12	40	15	15	9	1	13

5.3.3 Detailed SNP analysis on SSPs

SSPs in the reference genome upregulated post inoculation reported in Rouxel et al. (2011) were characterised for their SNP content in Zander (2015). Further detailed analysis on the SSPs is reported below.

5.3.3.1 PROVEAN prediction analysis

Of 21 SNPs listed in SSPs1-16 (Table 5.6), nine SNPs were synonymous mutations due to which there was a neutral effect of the SNP. Four SNPs were nonsynonymous but also neutral whereas three other nonsynonymous SNPs were predicted to be deleterious in nature. These deleterious SNPS had a PROVEAN score of less than -2.5 which is the default cut-off. Table 5.7 provides the annotation information for these SSPs. Most annotations pertained to housekeeping processes and the others were predicted proteins.

5.3.3.2 SSPs and RNA seq data

As per RNA-seq data available for isolate D5, the three most upregulated SSPs amongst the set of sixteen all contained SNPs in D5 (Table 5.6). No unique PAVs were reported for any of the SSPs in any of the isolates in this data set. Furthermore, none of the SSPs were lost in this data set.

Table 5.6 PROVEAN prediction of SNPs detected in SSPs and average expression in D5

(Highlighted pink-SNP mutations whose PROVEAN score is >-2.5 are predicted to be deleterious in nature; D-Deleterious mutation and N-neutral mutations; From/To-Standard abbreviations for amino acids; X-stop codon; Highlighted green-most expressed SSPs in this set; R-Reference; RNA seq-Average FPKM values for gene expression in D5)

Gene	SSP	SNP	R	D1	D2	D3	D4	D5	D6	D8	D9	D10	D12	RNA seq	From	To	Effect	Score
Lema_P003630.1	**SSP1**	A/C	C	C	C	C	C	A	C	C	C	C	C	132.97	L	L	N	0
		C/T	C	C	C	C	C	T	C	C	C	C	C	132.97	H	H	N	0
		A/G	A	A	A	A	A	G	A	A	A	A	A	132.97	I	V	N	-0.111
Lema_P013380.1	**SSP2**	G/T	G	T	T	G	T	G	T	T	G	G	T	26.58	Q	S	N	4
		A/G	A	G	G	A	G	A	G	G	A	A	G	26.58				
Lema_P119130.1	**SSP3**	A/G	G	A	A	G	A	G	A	G	G	G	G	29.56	S	F	D	-6
		C/T	T	C	C	T	C	T	C	T	T	T	T	29.56	L	L	N	0
Lema_P119620.1	**SSP4**	A/G	G	G	G	G	G	A	G	G	G	G	G	3.048	L	L	N	0
		C/T	C	C	C	C	T	C	C	C	C	C	C	3.048	P	L	D	-9.8333
Lema_P021120.1	**SSP5**	C/T	C	C	T	C	C	C	C	C	C	C	C	37.26	P	P	N	0
Lema_P028710.1	**SSP6**	A/G	A	A	A	A	G	A	A	A	A	A	A	403.44	T	T	N	0
Lema_P038290.1	**SSP7**	G/T	T	T	T	T	G	G	T	T	T	T	T	5.004	S	A	N	-0.8
Lema_P043210.1	**SSP8**	A/G	G	A	G	G	G	A	G	G	G	G	G	132.54	T	T	N	0
Lema_uP065940.1	**SSP9**	A/G	G	G	G	G	G	G	G	A	G	G	G	10.46	V	I	N	-1
Lema_P074890.1	**SSP10**	C/T	T	T	C	C	C	T	C	C	C	T	C	2.22	Q	Q	N	0
Lema_P075980.1	**SSP11**	A/C	C	C	C	C	A	C	C	C	C	C	C	36.95	V	V	N	0
Lema_P083940.1	**SSP12**	A/G	G	G	G	G	G	G	G	A	G	G	G	9.59	P	S	N	0.333
Lema_P093990.1	**SSP13**	A/G	A	G	G	G	G	G	G	A	G	A	A	8.056	P	P	N	0
Lema_P103890.1	**SSP14**	C/G	G	C	C	G	G	G	G	G	G	G	G	41.14	L	L	N	0
Lema_P106060.1	**SSP15**	A/T	A	A	A	T	T	T	T	T	A	A	T	9.04	P	P	N	0
Lema_P113200.1	**SSP16**	C/G	G	C	G	G	C	C	C	C	G	G	G	24.83	K	N	D	-4

Table 5.7 Annotation information for SSPs

Gene	SSP	Annotation
Lema_P003630.1	SSP1	Predicted protein
Lema_P013380.1	SSP2	Eukaryotic Translation Inititation Factor
Lema_P119130.1	SSP3	Predicted protein
Lema_P119620.1	SSP4	Lipase
Lema_P021120.1	SSP5	Gpi anchored protein
Lema_P028710.1	SSP6	Extracellular matrix protein
Lema_P038290.1	SSP7	Predicted protein
Lema_P043210.1	SSP8	Similar to expression library immunisation antigen-1
Lema_uP065940.1	SSP9	Predicted protein
Lema_P074890.1	SSP10	Pectate lyase
Lema_P075980.1	SSP11	Rhamnogalacturonan acetylesterase
Lema_P083940.1	SSP12	Predicted protein
Lema_P093990.1	SSP13	Similar to YesU protein
Lema_P103890.1	SSP14	Celp0028 effector like protein
Lema_P106060.1	SSP15	Predicted protein
Lema_P113200.1	SSP16	Predicted protein

5.4 Discussion

Results from PAVs, Gene loss and SNPs laid the structural brickwork for building a pan-genome in *L. maculans*. It enabled us to analyse the genomic content of blackleg from three different aspects. Most importantly, the combination of these resources provided us with a unique pipeline for the identification of novel *AvrLm* genes. We were also able to take into account and validate some of the results detailed in previous chapters of this thesis (Chapter 3 and Chapter 4).

5.4.1 Candidate effector genes

If the host plant harbours the corresponding resistance gene to the avirulence gene contained in a blackleg isolate, the pathogen is unable to cause infection. In such a scenario, the pathogen needs to overcome resistance and does so because of its high evolutionary potential. The detectable avirulence gene is quickly rendered unrecognisable through methods such as deletion, RIP mutation, point mutation or by conditionally losing the chromosome that contains the avirulence gene (Gout et al., 2006b; Fudal et al., 2007; Fudal et al., 2009; Balesdent et al., 2013; Ghanbarnia et al., 2014; Van de Wouw et al., 2014). This led us to hypothesise that the gene loss tool could highlight candidate effectors that are being lost in the sample set as compared to the reference. In the same token, PAV reads would map to unique genes present in one isolate versus the other and SNPs would highlight point mutations or RIP mutations in genes of interest.

5.4.1.1 Candidate A and B

As seen in Figure 5.4 and Figure 5.5, Candidate A is a candidate effector that has been successfully shortlisted keeping in mind the following attributes. It is located in an AT-rich region devoid of other predicted genes; it is 228 bp long and it is lost in D8 as compared to the reference. The first *AvrLm* gene that was characterised in *L. maculans* was *AvrLm1* (Gout et al., 2006b). It was thought to be "lost in the middle of nowhere" (Gout et al., 2006b), in a 296 Kb region rich in repeats. This candidate effector also shares similarities with other known avirulence genes. It is located in an AT-rich gene poor region. Usually, these gene-poor regions are riddled with repeat elements inactivated by RIP (Van de Wouw et al., 2014), as in the case for most known *AvrLm* genes. However, *AvrLm2* is an exception to this rule wherein it is located on a GC-rich island in the middle of an AT-rich region (Ghanbarnia et al., 2014). All other characterised *AvrLm* genes are cysteine-rich except for *AvrLm1*.

This gene is highly upregulated in planta when compared to other known *AvrLm* genes (Figure 5.12). The absence of this gene only in the isolate D8 does not follow any pattern of predicted *AvrLm* genes in the virulent phenotype data. However, based on previously mentioned attributes and the high ration of upregulation 7 dpi on susceptible plants indicates that the product of this gene is a probable effector or plays a significant role in the infection process. Further characterisation of this candidate will help demystify the protein structure and other attributes that may or may not be similar to currently known avirulence genes for this pathogen.

Candidate A was found to have 41% hit identity to Candidate B when tested on Blast2Go (Conesa et al., 2005). Candidate B, like Candidate A is also located in an AT-rich gene poor region on SC_15, affected by SNPs (Figure 5.5) and also shares other similarities to known effectors, based on which it was chosen as a candidate. Candidate B is also affected by deleterious SNP mutations, which are absent in D5. The interesting feature common to both these candidates is that their upregulation ratio in planta, 7 dpi is significantly high when compared to known avirulence gene expression, analysed in isolate D5 (Figure 5.12). This means that both these genes are important to the infection process in the pathogen. Lowe et al. (2014) also reported both these candidates as the top 20 upregulated genes 7 dpi and Candidate A in the top 20 upregulated 14 dpi, in the *L. maculans* isolate analysed.

The majority of effectors in fungal pathogen genomes are single genes and a rare few genes form small paralog families comprising up to three genes (de Guillen et al., 2015). These effectors do not share sequence homology with other effectors and also do not share any conserved motifs (de Guillen et al., 2015). Ohm et al. (2012) found that 21.3% of SSPs analysed in genomes of 18 Dothiodiomycetes did not have any homologues in any other genomes. Thus far, there have been no reports of fungal conserved motifs present across several effector sequences. There are however, some rare exceptions to the rule. The motif [YFW]xC is shared between a few rust fungi and effectors from *Blumeria graminis* f.sp. *hordei* (Godfrey et al., 2010; Duplessis et al., 2011). Ma et al. (2010) and Sperschneider et al. (2013) have stated that *Fusarium* species conserve a [SG]PC[KR]P motif. de Guillen et al. (2015) have recently discovered that effector genes belonging to distantly related species *Magnaporthe oryzae* and *Pyrenophora tritici-repentis* with no sequence similarity, share structural homology. These effectors have been grouped under the Magnaporthe Avr and Tox-B like-effector family (MAX). Other examples of fungal species that display structural conservation are *M. oryzae* (Zhang et al., 2013) and *Pyrenophora tritici-repentis* (Sarma et al., 2005) that all share β-sandwich structures. It is therefore largely unusual to notice sequence

similarity amongst genes that are involved in the infection process. The most recent discovery of *AvrLm3* (Plissonneau et al., 2015) revealed that it shared 29% identity at six conserved cysteine positions to the previously found *AvrLmJ1* (Van de Wouw et al., 2014). Interestingly, the alignment of candidates A and B protein products results in a conserved motif 'CLCTLxDTLWRM', lending the two genes 91% identity at the motif site (Figure 5.6). We also found 58% conservation at the motif site between *AvrLmJ1* and candidates A and B (Figure 5.7). This motif is also present in other isolates in our dataset. We deduce that this could be the first occurrence of a conserved motif amongst *L. maculans* isolates. Based on its conservation within this small group of samples and its high upregulation ration, we can hypothesise the following- It functions to provide the pathogen with an infection related ability. It might be related to gaining entry into the plant cell like the N-terminal motifs in *AvrL567* and *AvrM* (Rafiqi et al., 2010) and C-terminal motifs in *ToxA* effectors (Manning et al., 2008). It could also serve to translocate effectors into the host cell like RxLR-motifs in *Phytophthora* species (Jiang et al., 2008). Further study needs to be conducted to investigate its occurrence within population isolates and to functionally characterise the importance of the motif.

5.4.1.2 Candidate C

AvrLm5 is deemed to only be lost in isolate D5. Gene loss data and PAV data singled out Candidate C as a promising candidate for *AvrLm5*. This was also backed up by RNA-seq expression data that had no expression for Candidate C. It displays all the traits to be a potential effector including being present in an AT-rich region (Figure 5.8) with no other genes and does not share any homology with other know effectors. Further functional characterisation of this gene will help us in establishing its status as a true effector candidate.

5.4.1.3 SC_13 region of interest

As detailed in chapter 4, we discovered clustered gene loss on SC_13 in some isolates, which we attributed to high selection pressure acting in that region. (Zander, 2015) found evidence of similar selection pressure in the same region on SC_13 in their study. This theory is supported by the discovery of significant association of SNP73 (Figure 3.1, Chapter 3) to gene loss in D1, D3, D5 and D9 (See section 4.3.5, Chapter 4). Sequence analysis revealed that this SNP was located in a 1.7 kb region mutated by other SNPs (Table 5.2), the majority of which were G to A or C to T mutations. Therefore, we concluded that these SNPs were RIP mutations acting in the region immediately downstream of SC_13 gene loss. We also found that of ten predicted SNPs in that region, six SNPs, SNP1, 2, 4, 8, 9 and SNP10 were significantly associated with each other. This is

60% of the total SNPs analysed in the LD analysis that display significant p-values (Figure 5.10). RIP mutations have already been established as a fungal mechanism employed to initiate rapid evolution, usually brought on by stimulus like host recognition of *AvrLm* genes (Fudal et al., 2009; Grandaubert et al., 2014b; Rouxel et al., 2011), which is further evidence of high selection pressure in that region. Interestingly, this pattern of loss in specifically D1, D3, D5 and D9 also coincided with virulent phenotype on cultivars for *AvrLm8* (Chapter 4; See Table 4.5).

SNP73 was not predicted amongst the SNP identified within this data set (Table 5.2) because it was initially predicted based on two different parent isolates (Chapter 3). However, the initial results of low association, combined with the results from gene loss, helped us narrow down the region of interest. Also, amongst the 384 SNPs included in the Illumina GoldenGate assay, only 14 SNPs were included from SC_13. This number may have been too low to identify LD localised on SC_13 in a greater set of 384 SNPs.

This evidence of high selection pressure and probable open reading frames (Figure 5.9) indicated to us that there may be a gene of interest in this region. Although we were unable to predict signal peptides encoded in the region, there is a possibility that a novel, uncharacterised gene may be missing from the assembly of the reference genome for SC_13. Based on the pattern of loss in isolates D1, D3, D5 and D9, this novel gene could possibly be a candidate for *AvrLm8*. This type of occurrence has previously been identified in *C. tropicalis* where the gene *EFG1* which is responsible for biofilm formation, survival in the host and filamentation regulation (Mancera et al 2015) was initially reported to be missing. The gene was first identified in *C. albicans* and its ortholog was later identified to be located in a gap in the assembly of *C. tropicalis* (Mancera et al 2015). In lieu of these results, further inspection into these findings via sequencing and assembly of isolates used in this study, may shed more light on identifying the reason behind high selection pressure in that region.

5.4.1.4 AvrLm9

As per phenotypic studies, the reference genome v23.1.3 does not contain *AvrLm9*. Some avirulence genes have been predicted via linkage analysis to be clustered in the blackleg genome (Balesdent et al., 2002; Rouxel and Balesdent, 2005; Van de Wouw et al., 2010). These are the Avr1-2-6 and Avr3-4-7-9 clusters. Gout et al. (2006b), Fudal et al. (2007) and Ghanbarnia et al. (2014) accurately found *AvrLm1, AvrLm6* and *AvrLm2* to be clustered on SC_6 of the genome. *AvrLm4-7* has been found to be located on SC_12 (Parlange et al., 2009), as was *AvrLm3*

(Plissonneau et al., 2015). *AvrLm9* is yet to be characterised but we can hypothesise that they are clustered on SC_12 with *AvrLm4-7* and *AvrLm3*. In the instance that the pathogen utilises point mutations or RIP mutation to diversify *AvrLm9*, our SNP resource should highlight the same in the data. This means that *AvrLm9* could be annotated in sequence in the reference but mutations due to SNPs or RIP can cause it to be 'absent' as per the phenotype data. Based on these criteria, we narrowed down SNPs occurring in D1 vs other isolates to 30 SNPs of which one occurred within an annotated gene. However, that gene did not show other features that are typical of effector genes. This led us to believe that the *AvrLm9* method to confer virulence is a complete deletion of the gene and therefore we would not expect it to be present in the reference. Further analysis starting with de novo gene prediction using PAV comparisons might help identify this avirulence gene.

5.4.2 Genomics

Apart from utilising this amalgamated data resource for effector gene identification, we also gained some information about the genomics of the isolates in this dataset. Firstly, PAV reads for each isolate only mapped to 0.13-0.5% of predicted genes in the reference genome. This informs us of the vast genomic diversity between isolates in this species. From the available gene hits, we obtained the most hits to SC_1. SC_13 was also reported in Chapter 4 (See section 4.3.2, 4.3.5) to lose the second highest number of genes. According to the findings in Chapter 4, SC_13 may contain a probable secondary metabolite cluster and also houses a predicted effector gene candidate (Candidate D, section 5.3.1.2). We hypothesise that SC_13 is undergoing high selection pressure and contains unique, diverged genes, which in turn accounts for the high number of PAV gene hits. Other SCs including 2, 8, 9, 11, 12, 19 and 23 that do not show any PAV hits are deemed to contain genes that are common amongst this dataset, possibly housekeeping genes that would not be impacted by evolutionary forces acting on pathogenicity genes. PAV reads were also observed to map to intergenic regions due to those regions being less conserved (Figure 5.3).

As seen in Table 5.3, the PAV comparisons 'to' and 'vs' vary for each gene. Upon observing PAV read mappings to the genome, we noticed that reads map to different portions of the same gene for different isolates. PAV data is based on unique K-mer hits belonging to each individual isolate. K-mers are smaller parts of the whole sequence read that are highlighted as unique when a k-mer bit is present in one isolate versus absent in another. Once these K-mer hits are flagged, the whole sequence reads to which they belong to are flagged in turn and accumulated into a unique reads dataset for each isolate. When these reads are mapped to the genome, PAV reads mapping to gene annotations are known as 'gene hits' with a variation in its presence/absence in the dataset. PAVs not only identify present and absent genes based on sequence uniqueness but also differences that are emphasized due to highly diverged sequences. Therefore, it may be that if two isolates contain the same unique PAV gene hit but different 'PAV vs' isolates, that gene is diverged to different extents in different isolates. Also, what K-mer may be unique between one comparison may not be between another comparison because it may pick up different regions of the gene as the unique portion for each comparison.

Amongst the isolates analysed, D9 and D3 displayed the least PAV hits, whereas D1 and D6 displayed the most hits (Table 5.4). Hypothetically, this means that the former isolates share most similarities, most preserve their genomic content in this dataset and are under the least selection pressure. The opposite applies to the latter isolates where these isolates are the most unique,

different to other isolates, under most selection pressure leading to divergence and sequence difference. D9 and D3 were grouped together on the phylogenetic tree based on gene loss data (See section 4.3.5) as were D1 and D6. Despite these minor similarities, the sample size used here is too small to draw concrete conclusions about the evolutionary nature of these results. We can however, use this information to analyse the results from a larger population analysis.

Also seen in Table 5.3, the isolates in the 'PAV vs' column do not consistently correspond to the isolates in the 'gene loss' column. It is possible that in some cases, there may be low read coverage at that gene locus where no PAV reads are picked up compared to gene loss results. In other cases, we need to consider that K-mer that is picked up by the PAKAP method to conduct the uniqueness comparison. Between different comparisons, the method picks up different K-mer sizes of around 16-17 bp, for the same gene (Zander, 2015). Therefore, the 'PAV vs' results may not be consistent because different sections of the same gene can be lost or diverged in different isolates. Another possibility is that low read coverage for that gene accounts for it not being picked up in the isolate. Where the isolates are the same in the two columns, we can safely assume that the PAV reads highlight definite presence/absence on the basis of complete deletion in the 'vs' isolate.

De novo prediction combined with a pan-genome approach will enable us to gain an overall picture of the genomic content of this pathogen and highlight exact difference even within genes, between a core set of isolates.

5.4.3 SNPs in SSPs

SSPs explored by Rouxel et al. (2011) were sorted by Zander (2015) to analyse the SNP content of SSPs overexpressed 7 and 14 dpi in the reference genome (Table 5.5). We hypothesised that these sixteen SSPs were of significance to the pathogen due to the role them being overexpressed upon infection. This was supported by the gene loss data where none of the SSPs were reported as lost. SSPs containing synonymous and non-synonymous neutral mutations were supported by the RNA seq data in that the change didn't affect gene expression. On the other hand, the three SSPs predicted to contain deleterious SNPs, had low expression. However, we also found that even within this set of SSPs, low expression levels do not always correlate to the presence of a deleterious SNP mutation. This means that these genes are differently expressed in D5 as compared to the reference. Our previous findings may explain this. Thus far, we have found that the genomic content of each isolate is extremely diverse even within a small sample size. It was also exemplified by SNP analysis in a larger dataset in Chapter 3. Therefore, SSPs being overexpressed in the reference genome may not necessarily be expressed similarly in isolate D5. Functional redundancy may also account for low expression in SSPs perceived to be of fundamental importance to the pathogen. Further comparative studies of RNAseq data obtained from different isolates will help answer how these SSPs are expressed in different samples.

5.4.4 SSP annotations

We were also able to shed some light on putative functions of these SSPs (Table 5.6). SSP8 and SSP13 matched the annotation detailed in Rouxel et al. (2011). Most other SSPs were involved in housekeeping processes, such as SSP2 annotated as a eukaryotic translation initiation factor. Hypothetically, this SSP could be involved in one of eight stages of translation initiation which utilises nine translation initiation factors to assemble elongation component 80S-ribosomes (Jackson et al., 2010). However, eukaryotic translation initiation factors are not usually secreted. Therefore, the accuracy of gene annotation for this particular SSP needs to be investigated further. SSP4 was annotated a lipase is a hydrolase class enzyme that plays a catalytic role in hydrolysing long-chain triglycerides.

SSP5 is purported to be a GPI (glycosylphosphatidylinositol)-anchored protein, which are a type of lipid modification that occur on transmembrane proteins (Mayor and Riezman, 2004). In fungi, GPI-anchored proteins when attached, enables covalent transport of cell wall mannoproteins (Mayor and Riezman, 2004).

Extracellular matrix proteins as annotated for SSP6, in fungi, play a role in cell wall adhesion during plant infection (Ikeda et al., 2012). During infection, plant pathogenic fungi cover their surfaces with an extracellular matrix (ECM) which is composed of different compounds. This ECM enables the pathogen to better adhere to the plant surface during infection (Ikeda et al., 2012) and SSP6 product could be a probable component of the ECM.

SSP8 is predicted to code for an expression library immunisation (ELI) antigen-1. ELI was first developed to identify protective genes based on antigens encoded by invasive pathogen in their genomic sequence in order to create vaccines against those pathogens (Ivey et al., 2003). In terms of *L. maculans*, the sequence similarity to ELI antigen-1 may suggest that SSP8 plays an active role in host plant infection.

SSP10 annotated as pectate lyase is an enzyme that functions to degrade pectin-containing plant cell walls (Zhao et al., 2007). SSP11 too, encoding for rhamnogalacturonan acetylesterase is thought to play a role in cell wall degradation, as previously found when this category of genes was upregulated during appresorium development in *M. oryzae* (Soanes et al., 2012). Further work needs to be conducted to better functionally characterise SSP13 and SSP14.

5.4.5 Conclusion

When working with fungal pathogen genomes, it is important to address the gaps that cannot be filled with traditional bioinformatics approaches. This chapter exemplifies that the combination of different data resources obtained from bioinformatics tools, which model to identify a variety of genomic features, is a successful approach in genomic detailing. We were able to identify three effector gene candidates, of which two contain what may be the first conserved motif identified in this pathogen. Furthermore, we successfully validated our findings from Chapter 3 and Chapter 4 of high genomic diversity and occurrence of selection pressure in certain regions of the genome. This approach also allowed us to lay the groundwork for building a *L. maculans* pan-genome, which is sure to address any gaps in information amongst isolates of this species.

6 Thesis conclusion

When working with an elusive pathogen like *L. maculans*, every snippet of information about it proves invaluable to the *Brassica* community. However, exhaustive work conducted over the past few years has stripped away the intangibility of how this pathogen operates. This thesis contributes in decrypting another piece of the genomic puzzle of blackleg.

The major conclusion gleaned from combined work detailed here is the large-scale of genetic and genomic diversity amongst isolates of this fungus. Chapter 3 analyses population diversity amongst 69 different blackleg isolates using SNPs as molecular markers. The results showed an absence of any haplotype blocks associated with any parameters which was supported by linkage disequilibrium and principle component analyses. Based on this we concluded that the Australian blackleg population was panmictic where the members interacted randomly with one another via sexual reproduction. This led to a high degree of genetic diversity on a national scale. The fungus proliferates by spreading its sexual inoculum, ascospores through wind dispersal where ascopsores can travel several kilometres in the wind and human transport where humans unwittingly transport infected materials to different location. This enables varied genotypes of the pathogen to sexually reproduce, leading to further diversity and extensive recombination. McDonald and Linde (2002) described that a pathogen is afforded high evolutionary potential through aspects like maintaining a large population size, high genotype flow, mixed reproduction and high rate of mutation. Panmixia attributes the majority of these to the pathogen. Possessing a high evolutionary potential enables *L. maculans* to rapidly response to selection pressure imposed by the host.

Some evidence of localised clonality based on clusters of samples collected from the same state was also observed. The clusters were located across the phylogenetic tree, indicating genetic differences between clusters. We concluded that environmental conditions in each state would impact this pathogen differently. This theory needs to be explored in detail by conducting a large-scale population study comprising large samplings from each state. Our results were also indicative of high polymorphisms in the SNP dataset with PIC scores ranging from 0.3-0.5. We also validated the SNP prediction detailed in (Zander et al., 2013), which was the same SNP resource used to select the 384 SNPs used in this study. Work is already underway to initiate a large scale population study. Between 100-200 isolates for each year have been collected from a database of samples (Provided by the University of Melbourne and Marcroft Grains Pathology) isolated in the years

spanning 2009-2015. Next, the samples will be sequenced and further explored to build a pan-genome, a concept that will be discussed in detail later. Overall analysis of the dataset strongly supported panmixia.

Our results thus far urged us to explore the diversity of blackeg on a genomic level. The increasing popularity and availability of next-generation sequencing producing high-throughput data for small genomes made this possible. NGS data can be used in conjuction with a wide variety of bioinformatics tools. The results of similar analyses are detailed in chapters 4 and 5.

Our primary aim was to identify novel avirulence effectors that aid in causing infection on host plants. Fudal et al. (2009) reported that the major event conferring virulence on *Rlm1* cultivars was deletion of the *AvrLm1* locus. In *Rlm6* cultivars this was deletion and RIP mutations that highly diversified *AvrLm6*. In Chapter 4 we describe use of the bioinformatics tool SGSGeneLoss (Golicz et al., 2015) to identify genes lost from the genome of this pathogen, of which some could be candidate effectors. We also aimed to find which genes are necessary versus which genes are disposable to the fungus. Firstly, we were able to reach the same conclusions of high genetic diversity amongst isolates, as reported in Chapter. Genomic content was preserved and lost through no particular relation to any parameter, highlighting the variability within this species. An interesting result was isolate D4 losing a gene integral to the sirodesmin biosynthesis cluster, which were conserved by all other isolates. Sirodesmin is a fungal phytotoxin involved in lesion formation in stems colonisation. Preliminary infection studies did not differentiate D4 from the other isolates as being weaker in causing infections. However, characterising localised gene expression of the sirodesmin cluster genes in the stem might shed more light on the dysfunctionality of those genes in D4, via a stem infection assay.

A region under high selection pressure was also identified on SC_13, where gene loss was observed in blocks. Functional annotation of genes in the block revealed them to be a part of a probable secondary metabolite cluster. Secondary metabolites are identifiable by the presence of signature secondary metabolite genes like polyketide synthases, are disposable in nature and may be partially conserved within some genomes. The putative cluster we identified contained PKS1, was deleted in isolates D1, D3, D5 and D9 and conserved in other isolates. This set of genes needs to be further characterised to elucidate the functionality and the significance of the compound produced by the cluster.

We concluded that along with environmental factors, transposable element activity and RIP mutations also contribute to creating the genomic variation we observed in our results. Functionally annotating lost genes denoted them to be a part of a volatile group of genes whose absence in the genome does not afford a fitness cost to the pathogen. Furthermore, there was no observable loss of any one type of gene, signifying functional redundancy for lost genes.

Identifying novel effector candidates in pathogenic fungi like blackleg has always been a challenge due to no homology between known effectors or other effectors in related species. However, characterisation of other effector genes in blackleg has informed us that the pathogen deletes the locus containing the *AvrLm* gene (Fudal et al., 2007; Gout et al., 2006b) or it is mutated by SNPs (Ghanbarnia et al., 2014; Van de Wouw et al., 2014) or RIP affected to overcome host recognition (Parlange et al., 2009). With this in mind, gene loss data was combined with PAV data and SNP data from Zander (2015) to create a robust resource that was integrated into a pipeline for effector gene identification in blackleg. Chapter 5 established the details of the pipeline where the first step requires sequenced samples of interest using NGS. NGS data is then run through bioinformatics tools like SGSGeneLoss, SGSautoSNP (Lorenc et al., 2012) and the PAKAP method (Zander, 2015). Each tool provides us with a different picture of the blackleg genome, the combination of which allows for filling gaps that would otherwise be present when performing only one kind of analysis. The combined data resource is combed for candidate effectors highlighted as either deleted, highly varied and picked up by the PAV data or SNP mutated. These candidates must also display one or more properties common to blackleg effectors such as located in a gene sparse AT-rich region, upregulated during infection, cysteine rich and encoding SSPs. Finally, this was followed by Blast2Go (Conesa et al., 2005) and SignalP (Petersen et al., 2011) to determine functional annotations if any and probable signal peptides associated with the gene.

When employing this pipeline in blackleg, we were able to use reference genome annotations (Rouxel et al., 2011) to identify three novel candidates. Interestingly, candidates A and B shared 91% identity at a 12 aa motif conserved in the exons of the predicted gene. This motif was also 58.3% conserved when compared with *AvrLmJ1*. So far, there have been no reports of protein sequence conservation amongst any known blackleg effectors. This is the first report of motif conservation in this species and further work must be conducted to swiftly elucidate the functional significance of this motif. One probable gene was not predicted in the reference genome but significant association of SNPs in the region of interest suggest there may be a novel gene present. This can be investigated further via sequencing and assembly of isolates used in this study. All

three candidates warrant further characterisation via transient expression assays in tobacco with their *Rlm* gene counterparts. Another method would be to assess altered pathogenicity profiles on cultivars containing the corresponding *Rlm* gene after infecting with strain containing the candidate *AvrLm* gene with no functional copy of the gene.

Preliminary genomic analysis using this combined data resource validated the uniqueness of the putative secondary metabolite cluster predicted in Chapter 4. Furthermore, we were once again able to highlight that the genomic landscape of isolates of this pathogen, even when analysed in a small set, show high degrees of variation. Functional characterisation of SSPs which were reported as highly upregulated in the reference genome was also conducted. We concluded that these SSPs may be regulated differently in different isolates and RNA-seq expression analysis of these genes in a larger set of samples may be required. In this manner, we established the foundation of building a blackleg pan-genome. The penultimate aim and future work with this data will be to obtain NGS data and de novo assemble a core set of blackleg isolates whose characteristics hypothetically encompass major characteristics and genomic features of the species.

This thesis showed how various techniques can be utilised to analyse *L. maculans* at a population and genomic level. The three data resources used here exemplify how genomic resources are becoming increasingly useful in gaining more information about the biology of an organism. Furthermore, these tools are translational and can easily be employed to study the genomes of other organisms. Work should also be conducted in filling the gaps in annotation information of predicted genes in the blackleg genome. Obtaining annotation information will not only improve the genome but also better a vast variety of downstream data resources that use this information. Confirmed reports of high genetic variability within the population highlight the importance of continuing to decipher this pathogen. In the future, crucial next steps should include characterisation of these effector candidates. Investigating the molecular basis of the *AvrLm-Rlm* interaction should also be made a high priority however; work towards this has already commenced (Blondeau et al., 2015). The conserved motif identified in Chapter 5 must be characterised further for it may play a role in the infection process. Identifying disease-causing genes is a powerful tool in designing better disease-resistant cultivars that can ultimately be introduced in the field. Reduced disease occurrence leads to reduced crops losses and prolongs the viability of the canola industry in Australia. Ongoing national and international collaborations working with blackleg and canola are producing results that provide great reassurance to the optimistic future of the canola industry worldwide.

7 References

ABS. 2015. *Value of Agricultural Commodities Produced, Australia, year ended 30 June 2014* [Online]. Available: http://www.abs.gov.au/ausstats/abs@.nsf/mf/7503.0 [Accessed 8th December 2015].

ALBRECHTSEN, A., NIELSEN, F. C. & NIELSEN, R. 2010. Ascertainment biases in SNP chips affect measures of population divergence. *Molecular biology and Evolution,* 27, 2534-47.

AMBROSE, K. V., TIAN, Z., WANG, Y., SMITH, J., ZYLSTRA, G., HUANG, B. & BELANGER, F. C. 2015. Functional characterization of salicylate hydroxylase from the fungal endophyte *Epichloe festucae*. *Sci Rep,* 5, 10939.

AMSELEM, J., LEBRUN, M. H. & QUESNEVILLE, H. 2015. Whole genome comparative analysis of transposable elements provides new insight into mechanisms of their inactivation in fungal genomes. *BMC Genomics,* 16, 141.

ANSAN-MELAYAH, D., BALESDENT, M. H., DELOURME, R., PILET, M. L., TANGUY, X., RENARD, M. & ROUXEL, T. 1998. Genes for race-specific resistance against blackleg disease in *Brassica napus* L. *Plant Breeding,* 117, 373-378.

AOF. 2015. *Oilseeds Industry* [Online]. Australian Oilseeds Federation Inc. Available: http://www.australianoilseeds.com/oilseeds_industry. [Accessed 8th December 2015].

APPLEBY, N., EDWARDS, D. & BATLEY, J. 2009. New technologies for ultra-high throughput genotyping in plants. *In:* SOMERS, D. J., LANGRIDGE, P. & GUSTAFSON, J. P. (eds.) *Plant Genomics Methods and Protocols.* UK: Humana Press.

ARRACH, N., PORWOLLIK, S., CHENG, P., CHO, A., LONG, F., CHOI, S. H. & MCCLELLAND, M. 2008. *Salmonella serovar* identification using PCR-based detection of gene presence and absence. *Journal of Clinical Microbiology,* 46, 2581-2589.

ATTARD, A., GOUT, L., ROSS, S., PARLANGE, F., CATTOLICO, L., BALESDENT, M. H. & ROUXEL, T. 2005. Truncated and RIP-degenerated copies of the LTR retrotransposon Pholy are clustered in a pericentromeric region of the *Leptosphaeria maculans* genome. *Fungal Genetics and Biology,* 42, 30-41.

AYARPADIKANNAN, S. & KIM, H. S. 2014. The impact of transposable elements in genome evolution and genetic instability and their implications in various diseases. *Genomics and Informatics,* 12, 98-104.

BALESDENT, M. H., ATTARD, A., KUHN, M. L. & ROUXEL, T. 2002. New Avirulence Genes in the Phytopathogenic Fungus *Leptosphaeria maculans*. *Phytopathology,* 92, 1122-33.

BALESDENT, M. H., BARBETTI, M. J., LI, H., SIVASITHAMPARAM, K., GOUT, L. & ROUXEL, T. 2005. Analysis of *Leptosphaeria maculans* race structure in a worldwide collection of isolates. *Phytopathology,* 95, 1061-1071.

BALESDENT, M. H., FUDAL, I., OLLIVIER, B., BALLY, P., GRANDAUBERT, J., EBER, F., CHEVRE, A. M., LEFLON, M. & ROUXEL, T. 2013. The dispensable chromosome of *Leptosphaeria maculans* shelters an effector gene conferring avirulence towards *Brassica rapa*. *New Phytologist,* 198, 887-98.

BALL, A. D., STAPLEY, J., DAWSON, D. A., BIRKHEAD, T. R., BURKE, T. & SLATE, J. 2010. A comparison of SNPs and microsatellites as linkage mapping markers: lessons from the zebra finch (*Taeniopygia guttata*). *BMC Genomics,* 11, 218.

BARRINS, J. M., ADES, P. K., SALISBURY, P. A. & HOWLETT, B. J. 2004. Genetic diversity of Australian isolates of *Leptosphaeria maculans*, the fungus that causes blackleg of canola (*Brassica napus*). *Australasian Plant Pathology,* 33, 529-536.

BATLEY, J., BARKER, G., O'SULLIVAN, H., EDWARDS, K. J. & EDWARDS, D. 2003. Mining for single nucleotide polymorphisms and insertions/deletions in maize expressed sequence tag data. *Plant Physiology,* 132, 84-91.

BAUMANN, P., BAUMANN, L., LAI, C. Y., ROUHBAKHSH, D., MORAN, N. A. & CLARK, M. A. 1995. Genetics, physiology, and evolutionary relationships of the genus *Buchnera*: intracellular symbionts of aphids. *Annual Review of Microbiology,* 49, 55-94.
BEMIC COMPANY INC. 2010. Parafilm M.
BHATIA, Y., MISHRA, S. & BISARIA, V. S. 2002. Microbial beta-glucosidases: cloning, properties, and applications. *Critical Reviews in Biotechnology,* 22, 375-407.
BLONDEAU, K., BLAISE, F., GRAILLE, M., KALE, S. D., LINGLIN, J., OLLIVIER, B., LABARDE, A., LAZAR, N., DAVERDIN, G., BALESDENT, M. H., CHOI, D. H., TYLER, B. M., ROUXEL, T., VAN TILBEURGH, H. & FUDAL, I. 2015. Crystal structure of the effector *AvrLm4-7* of *Leptosphaeria maculans* reveals insights into its translocation into plant cell and recognition by resistance proteins. *The Plant Journal*, 83, 610-624.
BOKOR, A., BARBETTI, M. J., BROWN, A. G. P., MACNISH, G. C., POOLE, M. L. & WOOD, P. M. 1975. Blackleg-a major hazard to the rapeseed industry. *Journal of Agriculture, Western Australia*, 7-10.
BROWN, D. W., ADAMS, T. H. & KELLER, N. P. 1996. *Aspergillus* has distinct fatty acid synthases for primary and secondary metabolism. *Proc Natl Acad Sci U S A,* 93, 14873-7.
CAMBARERI, E. B., JENSEN, B. C., SCHABTACH, E. & SELKER, E. U. 1989. Repeat-Induced G-C to a-T Mutations in *Neurospora*. *Science,* 244, 1571-1575.
CAMPBELLS AUSTRALIA PTY LTD. 2015. *V8 Vegetable Juice* [Online]. Available: http://www.campbellsoup.com.au/brands-products/our-products/v8-vegetable-juice.aspx [Accessed 13th December 2015].
CELTON, J.-M., CHRISTOFFELS, A., SARGENT, D. J., XU, X. & REES, D. J. G. 2010. Genome-wide SNP identification by high-throughput sequencing and selective mapping allows sequence assembly positioning using a framework genetic linkage map. *BMC Biology,* 8, 1-8.
CHAGNÉ, D., BATLEY, J., EDWARDS, D. & FORSTER, J. W. 2007. Single Nucleotide Polymorphisms genotyping in plants. *In:* ORAGUZIE, N. C., RIKKERINK, E. H. A., GARDINER, S. E. & NIHAL DE SILVA, H. (eds.) *Association mapping in plants.* New York: Springer.
CHEN, S., SU, L., CHEN, J. & WU, J. 2013. Cutinase: Characteristics, preparation, and application. *Biotechnology Advances,* 31, 1754-1767.
CHOI, Y., SIMS, G. E., MURPHY, S., MILLER, J. R. & CHAN, A. P. 2012. Predicting the Functional Effect of Amino Acid Substitutions and Indels. *PloS One,* 7, 414-417.
CHUMA, I., ISOBE, C., HOTTA, Y., IBARAGI, K., FUTAMATA, N., KUSABA, M., YOSHIDA, K., TERAUCHI, R., FUJITA, Y., NAKAYASHIKI, H., VALENT, B. & TOSA, Y. 2011. Multiple Translocation of the *AVR-Pita* Effector Gene among Chromosomes of the Rice Blast Fungus *Magnaporthe oryzae* and Related Species. *PLoS Pathogens,* 7, e1002147.
CLAY, K. & SCHARDL, C. 2002. Evolutionary origins and ecological consequences of endophyte symbiosis with grasses. *The American Naturalist,* 160 Suppl 4, S99-S127.
CONESA, A., GÖTZ, S., GARCÍA-GÓMEZ, J. M., TEROL, J., TALÓN, M. & ROBLES, M. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics,* 21, 3674-3676.
COUCH, B. C., FUDAL, I., LEBRUN, M. H., THARREAU, D., VALENT, B., VAN KIM, P., NOTTEGHEM, J. L. & KOHN, L. M. 2005. Origins of host-specific populations of the blast pathogen *Magnaporthe oryzae* in crop domestication with subsequent expansion of pandemic clones on rice and weeds of rice. *Genetics,* 170, 613-630.
CUOMO, C. A., GUELDENER, U., XU, J. R., TRAIL, F., TURGEON, B. G., DI PIETRO, A., WALTON, J. D., MA, L. J., BAKER, S. E., REP, M., ADAM, G., ANTONIW, J., BALDWIN, T., CALVO, S., CHANG, Y. L., DECAPRIO, D., GALE, L. R., GNERRE, S., GOSWAMI, R. S., HAMMOND-KOSACK, K., HARRIS, L. J., HILBURN, K., KENNELL, J. C., KROKEN, S., MAGNUSON, J. K., MANNHAUPT, G., MAUCELI, E.,

MEWES, H. W., MITTERBAUER, R., MUEHLBAUER, G., MUNSTERKOTTER, M., NELSON, D., O'DONNELL, K., OUELLET, T., QI, W. H., QUESNEVILLE, H., RONCERO, M. I. G., SEONG, K. Y., TETKO, I. V., URBAN, M., WAALWIJK, C., WARD, T. J., YAO, J. Q., BIRREN, B. W. & KISTLER, H. C. 2007. The *Fusarium graminearum* genome reveals a link between localized polymorphism and pathogen specialization. *Science,* 317, 1400-1402.

DAI, Y., JIA, Y., CORRELL, J., WANG, X. & WANG, Y. 2010. Diversification and evolution of the avirulence gene *AVR-Pita1* in field isolates of *Magnaporthe oryzae*. *Fungal Genetics and Biology,* 47, 973-80.

DANGL, J. L. & JONES, J. D. G. 2001. Plant pathogens and integrated defence responses to infection. *Nature,* 411, 826-833.

DAVERDIN, G., ROUXEL, T., GOUT, L., AUBERTOT, J. N., FUDAL, I., MEYER, M., PARLANGE, F., CARPEZAT, J. & BALESDENT, M. H. 2012. Genome structure and reproductive behaviour influence the evolutionary potential of a fungal phytopathogen. *PLoS Pathogens,* 8, e1003020.

DE GUILLEN, K., ORTIZ-VALLEJO, D., GRACY, J., FOURNIER, E., KROJ, T. & PADILLA, A. 2015. Structure Analysis Uncovers a Highly Diverse but Structurally Conserved Effector Family in Phytopathogenic Fungi. *PLoS Pathogens,* 11, e1005228.

DE WIT, P. J., MEHRABI, R., VAN DEN BURG, H. A. and STERGIOPOULOS, I. 2009. Fungal effector proteins: past, present and future. *Molecular Plant Pathology*, 10, 735–747.

DILMAGHANI, A., BALESDENT, M. H., DIDIER, J. P., WU, C., DAVEY, J., BARBETTI, M. J., LI, H., MORENO-RICO, O., PHILLIPS, D., DESPEGHEL, J. P., VINCENOT, L., GOUT, L. & ROUXEL, T. 2009. The *Leptosphaeria maculans Leptosphaeria biglobosa* species complex in the American continent. *Plant Pathology,* 58, 1044-1058.

DILMAGHANI, A., GLADIEUX, P., GOUT, L., GIRAUD, T., BRUNNER, P. C., STACHOWIAK, A., BALESDENT, M. H. & ROUXEL, T. 2012. Migration patterns and changes in population biology associated with the worldwide spread of the oilseed rape pathogen *Leptosphaeria maculans*. *Molecular Ecology,* 21, 2519-2533.

DILMAGHANI, A., GOUT, L., MORENO-RICO, O., DIAS, J. S., COUDARD, L., CASTILLO-TORRES, N., BALESDENT, M. H. & ROUXEL, T. 2013. Clonal populations of *Leptosphaeria maculans* contaminating cabbage in Mexico. *Plant Pathology,* 62, 520-532.

DODDS, P. N. & RATHJEN, J. P. 2010. Plant immunity: towards an integrated view of plant-pathogen interactions. *Nature Reviews Genetics,* 11, 539-48.

DRAY, S. & DUFOUR, A. B. 2007. The ade4 package: implementing the duality diagram for ecologists. *Journal of Statistical Software,* 22, 1-20.

DUPLESSIS, S., CUOMO, C. A., LIN, Y. C., AERTS, A., TISSERANT, E., VENEAULT-FOURREY, C., JOLY, D. L., HACQUARD, S., AMSELEM, J., CANTAREL, B. L., CHIU, R., COUTINHO, P. M., FEAU, N., FIELD, M., FREY, P., GELHAYE, E., GOLDBERG, J., GRABHERR, M. G., KODIRA, C. D., KOHLER, A., KUES, U., LINDQUIST, E. A., LUCAS, S. M., MAGO, R., MAUCELI, E., MORIN, E., MURAT, C., PANGILINAN, J. L., PARK, R., PEARSON, M., QUESNEVILLE, H., ROUHIER, N., SAKTHIKUMAR, S., SALAMOV, A. A., SCHMUTZ, J., SELLES, B., SHAPIRO, H., TANGUAY, P., TUSKAN, G. A., HENRISSAT, B., VAN DE PEER, Y., ROUZE, P., ELLIS, J. G., DODDS, P. N., SCHEIN, J. E., ZHONG, S., HAMELIN, R. C., GRIGORIEV, I. V., SZABO, L. J. & MARTIN, F. 2011. Obligate biotrophy features unraveled by the genomic analysis of rust fungi. *Proceedings of the National Academy of Sciences of the United States of America,* 108, 9166-71.

DURAN, C., BOSKOVIC, Z., IMELFORT, M., BATLEY, J., HAMILTON, N. A. & EDWARDS, D. 2010. CMap3D: a 3D visualization tool for comparative genetic maps. *Bioinformatics,* 26, 273-274.

DURSTEWITZ, G., POLLEY, A., PLIESKE, J., LUERSSEN, H., GRANER, E. M., WIESEKE, R. & GANAL, M. W. 2010. SNP discovery by amplicon sequencing and multiplex SNP genotyping in the allopolyploid species *Brassica napus*. *Genome,* 53, 948-956.
EDWARDS, D. & BATLEY, J. 2004. Plant bioinformatics: from genome to phenome. *Trends in Biotechnology,* 22, 232-237.
EDWARDS, D., BATLEY, J., COGAN, N. O. I., FORSTER, J. W. & CHAGNÉ, D. 2007. Single Nucleotide Polymorphism discovery. *In:* ORAGUZIE, N. C., RIKKERINK, E. H. A., GARDINER, S. E. & NIHAL DE SILVA, H. (eds.) *Association mapping in plants.* New York: Springer.
EDWARDS, D., BATLEY, J. & SNOWDON, R. J. 2013. Accessing complex crop genomes with next-generation sequencing. *Theoretical and Applied Genetics,* 126, 1-11.
ELLIOTT, C. E., CALLAHAN, D. L., SCHWENK, D., NETT, M., HOFFMEISTER, D. & HOWLETT, B. J. 2013. A gene cluster responsible for biosynthesis of phomenoic acid in the plant pathogenic fungus, *Leptosphaeria maculans*. *Fungal Genetics and Biology,* 53, 50-58.
ELLIOTT, C. E., GARDINER, D. M., THOMAS, G., COZIJNSEN, A., A, V. D. W. & HOWLETT, B. J. 2007. Production of the toxin sirodesmin PL by *Leptosphaeria maculans* during infection of *Brassica napus*. *Molecular Plant Pathology,* 8, 791-802.
ELLIOTT, C. E., HARJONO & HOWLETT, B. J. 2008. Mutation of a gene in the fungus *Leptosphaeria maculans* allows increased frequency of penetration of stomatal apertures of Arabidopsis thaliana. *Molecular Plant Pathology,* 1, 471-81.
ELLIOTT, C. E. & HOWLETT, B. J. 2006. Overexpression of a 3-ketoacyl-CoA thiolase in *Leptosphaeria maculans* causes reduced pathogenicity on *Brassica napus*. *Molecular Plant-Microbe Interactions,* 19, 588-96.
ELLIS, J. G., RAFIQI, M., GAN, P., CHAKRABARTI, A. & DODDS, P. N. 2009. Recent progress in discovery and functional analysis of effector proteins of fungal and oomycete plant pathogens. *Current Opinion in Plant Biology,* 12, 399-405.
FENG, J., ZHANG, H., STRELKOV, S. E. & HWANG, S. F. 2014. The *LmSNF1* Gene Is Required for Pathogenicity in the Canola Blackleg Pathogen *Leptosphaeria maculans*. *PloS One,* 9, e92503.
FLOR, H. H. 1942. Inheritance of pathogenicity in Melampsora lini. *Phytopathology,* 32, 653-669.
FOWLER, T. & BROWN, R. D. 1992. The *Bgl1* Gene Encoding Extracellular Beta-Glucosidase from *Trichoderma Reesei* Is Required for Rapid Induction of the Cellulase Complex. *Molecular Microbiology,* 6, 3225-3235.
FREITAG, M., WILLIAMS, R. L., KOTHE, G. O. & SELKER, E. U. 2002. A cytosine methyltransferase homologue is essential for repeat-induced point mutation in *Neurospora crassa*. *Proceedings of the National Academy of Sciences of the United States of America,* 99, 8802-8807.
FUDAL, I., ROSS, S., BRUN, H., BESNARD, A.-L., ERMEL, M., KUHN, M.-L., BALESDENT, M.-H. & ROUXEL, T. 2009. Repeat-Induced Point Mutation (RIP) as an Alternative Mechanism of Evolution Toward Virulence in *Leptosphaeria maculans*. *Molecular Plant-Microbe Interactions,* 22, 932-941.
FUDAL, I., ROSS, S., GOUT, L., BLAISE, F., KUHN, M. L., ECKERT, M. R., CATTOLICO, L., BERNARD-SAMAIN, S., BALESDENT, M. H. & ROUXEL, T. 2007. Heterochromatin-like regions as ecological niches for avirulence genes in the *Leptosphaeria maculans* genome: Map-based cloning of *AvrLm6*. *Molecular Plant-Microbe Interactions,* 20, 459-470.
FULTON, T. M., CHUNWONGSE, J. & TANKSLEY, S. D. 1995. Microprep Protocol for Extraction of DNA from Tomato and Other Herbaceous Plants. *Plant Molecular Biology Reporter,* 13, 207-209.

GALAGAN, J. E., CALVO, S. E., BORKOVICH, K. A., SELKER, E. U., READ, N. D., JAFFE, D., FITZHUGH, W., MA, L. J., SMIRNOV, S., PURCELL, S., REHMAN, B., ELKINS, T., ENGELS, R., WANG, S. G., NIELSEN, C. B., BUTLER, J., ENDRIZZI, M., QUI, D. Y., IANAKIEV, P., PEDERSEN, D. B., NELSON, M. A., WERNER-WASHBURNE, M., SELITRENNIKOFF, C. P., KINSEY, J. A., BRAUN, E. L., ZELTER, A., SCHULTE, U., KOTHE, G. O., JEDD, G., MEWES, W., STABEN, C., MARCOTTE, E., GREENBERG, D., ROY, A., FOLEY, K., NAYLOR, J., STABGE-THOMANN, N., BARRETT, R., GNERRE, S., KAMAL, M., KAMVYSSELIS, M., MAUCELI, E., BIELKE, C., RUDD, S., FRISHMAN, D., KRYSTOFOVA, S., RASMUSSEN, C., METZENBERG, R. L., PERKINS, D. D., KROKEN, S., COGONI, C., MACINO, G., CATCHESIDE, D., LI, W. X., PRATT, R. J., OSMANI, S. A., DESOUZA, C. P. C., GLASS, L., ORBACH, M. J., BERGLUND, J. A., VOELKER, R., YARDEN, O., PLAMANN, M., SEILER, S., DUNLAP, J., RADFORD, A., ARAMAYO, R., NATVIG, D. O., ALEX, L. A., MANNHAUPT, G., EBBOLE, D. J., FREITAG, M., PAULSEN, I., SACHS, M. S., LANDER, E. S., NUSBAUM, C. & BIRREN, B. 2003. The genome sequence of the filamentous fungus *Neurospora crassa*. *Nature,* 422, 859-868.

GALAGAN, J. E. & SELKER, E. U. 2004. RIP: the evolutionary cost of genome defense. *Trends in Genetics,* 20, 417-423.

GARDINER, D. M., COZIJNSEN, A. J., WILSON, L. M., PEDRAS, M. S. & HOWLETT, B. J. 2004. The sirodesmin biosynthetic gene cluster of the plant pathogenic fungus Leptosphaeria maculans. *Molecular Microbiology,* 53, 1307-18.

GE HEALTHCARE LIFE SCIENCES 2015. Whatman, Grade AA.

GHANBARNIA, K., FUDAL, I., LARKAN, N. J., LINKS, M. G., BALESDENT, M. H., PROFOTOVA, B., FERNANDO, W. G., ROUXEL, T. & BORHAN, M. H. 2014. Rapid identification of the *Leptosphaeria maculans* avirulence gene *AvrLm2* using an intraspecific comparative genomics approach. *Molecular Plant Pathology*.

GIRAUD, T., ENJALBERT, J., FOURNIER, E., DELMOTTE, F. & DUTECH, C. 2008. Population genetics of fungal diseases of plants. *Parasite-Journal De La Societe Francaise De Parasitologie,* 15, 449-454.

GLADIEUX, P., ROPARS, J., BADOUIN, H., BRANCA, A., AGUILETA, G., DE VIENNE, D. M., DE LA VEGA, R. C. R., BRANCO, S. & GIRAUD, T. 2014. Fungal evolutionary genomics provides insight into the mechanisms of adaptive divergence in eukaryotes. *Molecular Ecology,* 23, 753-773.

GODFREY, D., BOHLENIUS, H., PEDERSEN, C., ZHANG, Z., EMMERSEN, J. & THORDAL-CHRISTENSEN, H. 2010. Powdery mildew fungal effector candidates share N-terminal Y/F/WxC-motif. *BMC Genomics,* 11, 317.

GOLICZ, A. A., MARTINEZ, P. A., ZANDER, M., PATEL, D. A., VAN DE WOUW, A. P., VISENDI, P., FITZGERALD, T. L., EDWARDS, D. & BATLEY, J. 2015. Gene loss in the fungal canola pathogen *Leptosphaeria maculans*. *Functional and Integrative Genomics,* 15, 189-96.

GONZALEZ, V. M., AVENTIN, N., CENTENO, E. & PUIGDOMENECH, P. 2013. High presence/absence gene variability in defense-related gene clusters of *Cucumis melo*. *BMC Genomics,* 14, 782.

GOUT, L., ECKERT, M., ROUXEL, T. & BALESDENT, M. H. 2006a. Genetic variability and distribution of mating type alleles in field populations of *Leptosphaeria maculans* from France. *Applied Environmental Microbiology,* 72, 185-91.

GOUT, L., FUDAL, I., KUHN, M. L., BLAISE, F., ECKERT, M., CATTOLICO, L., BALESDENT, M. H. & ROUXEL, T. 2006b. Lost in the middle of nowhere: the AvrLm1 avirulence gene of the Dothideomycete *Leptosphaeria maculans*. *Molecular Microbiology,* 60, 67-80.

GRANDAUBERT, J., BALESDENT, M.-H. & ROUXEL, T. 2014a. Evolutionary and Adaptive Role of Transposable Elements in Fungal Genomes. *In:* FRANCIS, M. M. (ed.) *Advances in Botanical Research.* Academic Press.

GRANDAUBERT, J., LOWE, R. G., SOYER, J. L., SCHOCH, C. L., VAN DE WOUW, A. P., FUDAL, I., ROBBERTSE, B., LAPALU, N., LINKS, M. G., OLLIVIER, B., LINGLIN, J., BARBE, V., MANGENOT, S., CRUAUD, C., BORHAN, H., HOWLETT, B. J., BALESDENT, M. H. & ROUXEL, T. 2014b. Transposable element-assisted evolution and adaptation to host plant within the *Leptosphaeria maculans*-L*eptosphaeria biglobosa* species complex of fungal pathogens. *BMC Genomics,* 15, 891.

HANE, J. K., LOWE, R. G. T., SOLOMON, P. S., TAN, K. C., SCHOCH, C. L., SPATAFORA, J. W., CROUS, P. W., KODIRA, C., BIRREN, B. W., GALAGAN, J. E., TORRIANI, S. F. F., MCDONALD, B. A. & OLIVER, R. P. 2007. Dothideomycete-plant interactions illuminated by genome sequencing and EST analysis of the wheat pathogen Stagonospora nodorum. *Plant Cell,* 19, 3347-3368.

HANE, J. K., WILLIAMS, A. H., TARANTO, A. P., SOLOMON, P. S. & OLIVER, R. P. 2015. Repeat-Induced Point Mutation: A Fungal-Specific, Endogenous Mutagenesis Process. *In*: VAN DER BERG, M. A. & MARUTHACHALAM, K. (eds.) *Genetic Transformation Systems in Fungi, Vol 2.* Springer International Publishing.

HAYDEN, H. L., COZIJNSEN, A. J. & HOWLETT, B. J. 2007. Microsatellite and minisatellite analysis of *Leptosphaeria maculans* in Australia reveals regional genetic differentiation. *Phytopathology,* 97, 879-887.

HAYWARD, A., MCLANDERS, J., CAMPBELL, E., EDWARDS, D. & BATLEY, J. 2012. Genomic advances will herald new insights into the *Brassica*: *Leptosphaeria maculans* pathosystem. *Plant Biology,* 14, 1-10.

HOGENHOUT, S. A., VAN DER HOORN, R. A. L., TERAUCHI, R. & KAMOUN, S. 2009. Emerging Concepts in Effector Biology of Plant-Associated Organisms. *Molecular Plant-Microbe Interactions,* 22, 115-122.

HOLDEN, S. 2012. *Growing Canola* [Online]. Agriculture Victoria: Department of Environment and Primary Industries. Available: http://agriculture.vic.gov.au/agriculture/grains-and-other-crops/crop-production/growing-canola [Accessed 8th December 2015].

HOWLETT, B. J., LOWE, R. G. T., MARCROFT, S. J. & VAN DE WOUW, A. P. 2015. Evolution of virulence in fungal plant pathogens: exploiting fungal genomics to control plant disease. *Mycologia,* 107, 441-451.

HUA-VAN, A., LE ROUZIC, A., BOUTIN, T. S., FILEE, J. & CAPY, P. 2011. The struggle for life of the genome's selfish architects. *Biology Direct,* 6, 19.

HUANG, J., SI, W. N., DENG, Q. M., LI, P. & YANG, S. H. 2014. Rapid evolution of avirulence genes in rice blast fungus *Magnaporthe oryzae*. *BMC Genetics,* 15.

IAFRATE, A. J., FEUK, L., RIVERA, M. N., LISTEWNIK, M. L., DONAHOE, P. K., QI, Y., SCHERER, S. W. & LEE, C. 2004. Detection of large-scale variation in the human genome. *Nature Genetics,* 36, 949-51.

ICHINOSE, Y., TAGUCHI, F. & MUKAIHARA, T. 2013. Pathogenicity and virulence factors of *Pseudomonas syringae*. *Journal of General Plant Pathology,* 79, 285-296.

IDNURM, A. & HOWLETT, B. J. 2003. Analysis of loss of pathogenicity mutants reveals that repeat-induced point mutations can occur in the Dothideomycete *Leptosphaeria maculans*. *Fungal Genetics and Biology*. 39, 31-37.

IKEDA, K., INOUE, K., KITAGAWA, H., MEGURO, H., SHIMOI, S. & PARK, P. 2012. The Role of the Extracellular Matrix (ECM) in Phytopathogenic Fungi: A Potential Target for Disease Control. [Online]. Available: http://www.intechopen.com/books/howtoreference/plant-pathology/the-role-of-the-extracellular-matrix-ecm-in-phytopathogenic-fungi-a-potential-target-for-disease-con. [Accessed 13th December 2015].

ILLUMINA INC. 2015. [Online]. Available: http://www.illumina.com/ [Accessed 13th December 2015].
ILLUMINA INC. 2013. *GenomeStudio software* [Online]. Available: http://www.illumina.com/software/genomestudio_software.ilmn. [Accessed 8th December 2015].
IMELFORT, M., DURAN, C., BATLEY, J. & EDWARDS, D. 2009. Discovering genetic polymorphisms in next-generation sequencing data. *Plant Biotechnology Journal,* 7, 312-317.
IVEY, F. D., MAGEE, D. M., WOITASKE, M. D., JOHNSTON, S. A. & COX, R. A. 2003. Identification of a protective antigen of *Coccidioides immitis* by expression library immunization. *Vaccine,* 21, 4359-67.
JACKSON, R. J., HELLEN, C. U. & PESTOVA, T. V. 2010. The mechanism of eukaryotic translation initiation and principles of its regulation. *Nature Reviews Molecular Cell Biology,* 11, 113-27.
JIANG, R. H. Y., TRIPATHY, S., GOVERS, F. & TYLER, B. M. 2008. RXLR effector reservoir in two *Phytophthora* species is dominated by a single rapidly evolving superfamily with more than 700 members. *Proceedings of the National Academy of Sciences of the United States of America,* 105, 4874-4879.
JURKOWSKI, T. P. & JELTSCH, A. 2011. On the evolutionary origin of eukaryotic DNA methyltransferases and Dnmt2. *PLoS One,* 6, e28104.
KATINKA, M. D., DUPRAT, S., CORNILLOT, E., METENIER, G., THOMARAT, F., PRENSIER, G., BARBE, V., PEYRETAILLADE, E., BROTTIER, P., WINCKER, P., DELBAC, F., EL ALAOUI, H., PEYRET, P., SAURIN, W., GOUY, M., WEISSENBACH, J. & VIVARES, C. P. 2001. Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. *Nature,* 414, 450-3.
KAUR, S., COGAN, N. O. I., YE, G., BAILLIE, R. C., HAND, M. L., LING, A. E., MCGEAREY, A. K., KAUR, J., HOPKINS, C. J., TODOROVIC, M., MOUNTFORD, H., EDWARDS, D., BATLEY, J., BURTON, W., SALISBURY, P., GORORO, N., MARCROFT, S., KEARNEY, G., SMITH, K. F., FORSTER, J. W. & SPANGENBERG, G. C. 2009. Genetic map construction and QTL mapping of resistance to blackleg (*Leptosphaeria maculans*) disease in Australian canola (*Brassica napus* L.) cultivars. *Theoretical and Applied Genetics,* 120, 71-83.
KEARSE, M., MOIR, R., WILSON, A., STONES-HAVAS, S., CHEUNG, M., STURROCK, S., BUXTON, S., COOPER, A., MARKOWITZ, S., DURAN, C., THIERER, T., ASHTON, B., MEINTJES, P. & DRUMMOND, A. 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics,* 28, 1647-9.
KELKAR, H. S., GRIFFITH, J., CASE, M. E., COVERT, S. F., HALL, R. D., KEITH, C. H., OLIVER, J. S., ORBACH, M. J., SACHS, M. S., WAGNER, J. R., WEISE, M. J., WUNDERLICH, J. K. & ARNOLD, J. 2001. The *Neurospora crassa* genome: cosmid libraries sorted by chromosome. *Genetics,* 157, 979-90.
KELLER, N. P., TURNER, G. & BENNETT, J. W. 2005. Fungal secondary metabolism - from biochemistry to genomics. *Nature Reviews Microbiology,* 3, 937-47.
KHANGURA R.K., BARBETTI M.J. 2004. Time of sowing and fungicides affect blackleg (*Leptosphaeria maculans*) severity and yield in canola. *Australian Journal of Experimental Agriculture*, 44, 1205–1213.
KRYLOV, D. M., WOLF, Y. I., ROGOZIN, I. B. & KOONIN, E. V. 2003. Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. *Genome Research,* 13, 2229-2235.

KUO, A., BUSHNELL, B. & GRIGORIEV, I. V. 2014. Chapter One - Fungal Genomics: Sequencing and Annotation. *In:* FRANCIS, M. M. (ed.) *Advances in Botanical Research.* Academic Press.

LEWIN-KOH, N. J., BIVAND, R., PEBESMA, E. J., ARCHER, E., BADDELEY, A., BIBIKO, H., CALLAHAN, J., CARRILLO, G., DRAY, S., FORREST, D., FRIENDLY, M., GIRAUDOUX, P., GOLICHER, D., RUBIO, V. G., HAUSMANN, P., HUFTHAMMER, K. O., JAGGER, T., LUQUE, S. P., MACQUEEN, D., NICCOLAI, A., LAMIGUEIRO, O. P., SHORT, T., SNOW, G., STABLER, B. & TURNER, R. 2012. maptools: Tools for reading and handling spatial objects. R package version 0.8-20 ed.

LI, R. Q., YU, C., LI, Y. R., LAM, T. W., YIU, S. M., KRISTIANSEN, K. & WANG, J. 2009. SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics,* 25, 1966-1967.

LIFE TECHNOLOGIES 2015. Qubit Flurometer.

LIM, F. Y., SANCHEZ, J. F., WANG, C. C. & KELLER, N. P. 2012. Toward awakening cryptic secondary metabolite gene clusters in filamentous fungi. *Methods in Enzymology,* 517, 303-24.

LIU, N. J., CHEN, L., WANG, S., OH, C. G. & ZHAO, H. Y. 2005. Comparison of single-nucleotide polymorphisms and microsatellites in inference of population structure. *Bmc Genetics,* 6.

LORENC, M. T., HAYASHI, S., STILLER, J., HONG, L., MANOLI, S., RUPERAO, P., VISENDI, P., BERKMAN, P. J., LAI, K., BATLEY, J. & EDWARDS, D. 2012. Discovery of Single Nucleotide Polymorphisms in Complex Genomes Using SGSautoSNP. *Biology,* 1, 370-382.

LOWE, R. G., CASSIN, A., GRANDAUBERT, J., CLARK, B. L., VAN DE WOUW, A. P., ROUXEL, T. & HOWLETT, B. J. 2014. Genomes and transcriptomes of partners in plant-fungal-interactions between canola (*Brassica napus*) and two *Leptosphaeria* species. *PLoS One,* 9, e103098.

MA, L. J., VAN DER DOES, H. C., BORKOVICH, K. A., COLEMAN, J. J., DABOUSSI, M. J., DI PIETRO, A., DUFRESNE, M., FREITAG, M., GRABHERR, M., HENRISSAT, B., HOUTERMAN, P. M., KANG, S., SHIM, W. B., WOLOSHUK, C., XIE, X., XU, J. R., ANTONIW, J., BAKER, S. E., BLUHM, B. H., BREAKSPEAR, A., BROWN, D. W., BUTCHKO, R. A., CHAPMAN, S., COULSON, R., COUTINHO, P. M., DANCHIN, E. G., DIENER, A., GALE, L. R., GARDINER, D. M., GOFF, S., HAMMOND-KOSACK, K. E., HILBURN, K., HUA-VAN, A., JONKERS, W., KAZAN, K., KODIRA, C. D., KOEHRSEN, M., KUMAR, L., LEE, Y. H., LI, L., MANNERS, J. M., MIRANDA-SAAVEDRA, D., MUKHERJEE, M., PARK, G., PARK, J., PARK, S. Y., PROCTOR, R. H., REGEV, A., RUIZ-ROLDAN, M. C., SAIN, D., SAKTHIKUMAR, S., SYKES, S., SCHWARTZ, D. C., TURGEON, B. G., WAPINSKI, I., YODER, O., YOUNG, S., ZENG, Q., ZHOU, S., GALAGAN, J., CUOMO, C. A., KISTLER, H. C. & REP, M. 2010. Comparative genomics reveals mobile pathogenicity chromosomes in *Fusarium*. *Nature,* 464, 367-73.

MAJOR SCIENCE. 2015. *SmartView Pro Imager System* [Online]. Available: http://www.majorsci.com/4-002-UVCI1000series.html [Accessed 13th December 2015].

MANCERA E., PORMAN, A. M., CUOMO, C. A., BENNETT, A. J. & JOHNSON, A.D. 2015. Finding a Missing Gene: *EFG1* Regulates Morphogenesis in *Candida tropicalis*. *G3 (Bethesda)*. 5, 849–856.

MANNING, V. A., HAMILTON, S. M., KARPLUS, P. A. & CIUFFETTI, L. M. 2008. The Arg-Gly-Asp-containing, solvent-exposed loop of Ptr ToxA is required for internalization. *Molecular Plant-Microbe Interactions,* 21, 315-325.

MARCROFT, S. & BLUETT, C. 2008. Blackleg of canola. *Agriculture notes, Department of Primary Industries* [Online], AG1352. Available:

http://www.dpi.vic.gov.au/agriculture/pests-diseases-and-weeds/plant-diseases/grains-pulses-cereals/blackleg-of-canola. [Accessed 15th December 2015].
MARCROFT, S. J., ELLIOTT, V. L., COZIJNSEN, A. J., SALISBURY, P. A., HOWLETT, B. J. & VAN DE WOUW, A. P. 2012. Identifying resistance genes to *Leptosphaeria maculans* in Australian *Brassica napus* cultivars based on reactions to isolates with known avirulence genotypes. *Crop & Pasture Science,* 63, 338-350.
MARCROFT, S. J., SPRAGUE, S. J., PYMER, S. J., SALISBURY, P. A. & HOWLETT, B. J. 2003. Factors affecting production of inoculum of the blackleg fungus (*Leptosphaeria maculans*) in south-eastern Australia. *Australian Journal of Experimental Agriculture,* 43, 1231-1236.
MAYOR, S. & RIEZMAN, H. 2004. Sorting GPI-anchored proteins. *Nature Reviews Molecular Cell Biology,* 5, 110-20.
MCDONALD, B. A. & LINDE, C. 2002. Pathogen population genetics, evolutionary potential, and durable resistance. *Annual Review of Phytopathology,* 40, 349-+.
MCHALE, L. K., HAUN, W. J., XU, W. W., BHASKAR, P. B., ANDERSON, J. E., HYTEN, D. L., GERHARDT, D. J., JEDDELOH, J. A. & STUPAR, R. M. 2012. Structural variants in the soybean genome localize to clusters of biotic stress-response genes. *Plant Physiology,* 159, 1295-308.
MCVEAN, G. 2008. Linkage Disequilibrium, Recombination and Selection. *In:* BALDING, D. J., BISHOP, M. & CANNINGS, C. (eds.) *Handbook of Statistical Genetics.* Third ed. Chichester, UK: John Wiley & Sons, Ltd.
MEHRABI, R., BAHKALI, A. H., ABD-ELSALAM, K. A., MOSLEM, M., BEN M'BAREK, S., GOHARI, A. M., JASHNI, M. K., STERGIOPOULOS, I., KEMA, G. H. & DE WIT, P. J. 2011. Horizontal gene and chromosome transfer in plant pathogenic fungi affecting host range. *FEMS Microbiology Reviews,* 35, 542-54.
MERCK MILLIPORE CORPORATION 2015. Milli-Q.
MISHRA, S., RAO, S. & DEB, J. K. 1989. Isolation and Characterization of a Mutant of *Trichoderma reesei* Showing Reduced Levels of Extracellular Beta-Glucosidase. *Journal of General Microbiology,* 135, 3459-3465.
MURRAY, G. M. & BRENNAN, J. P. 2012. The current and potential costs from diseases of oilseed crops in australia. *Grains Research and Development Corporation* [Online]. Available: https://grdc.com.au/Resources/Publications/2012/06/The-Current-and-Potential-Costs-from-Diseases-of-Oilseed-Crops-in-Australia. [Accessed 13th December 2015].
OHM, R. A., FEAU, N., HENRISSAT, B., SCHOCH, C. L., HORWITZ, B. A., BARRY, K. W., CONDON, B. J., COPELAND, A. C., DHILLON, B., GLASER, F., HESSE, C. N., KOSTI, I., LABUTTI, K., LINDQUIST, E. A., LUCAS, S., SALAMOV, A. A., BRADSHAW, R. E., CIUFFETTI, L., HAMELIN, R. C., KEMA, G. H., LAWRENCE, C., SCOTT, J. A., SPATAFORA, J. W., TURGEON, B. G., DE WIT, P. J., ZHONG, S., GOODWIN, S. B. & GRIGORIEV, I. V. 2012. Diverse lifestyles and strategies of plant pathogenesis encoded in the genomes of eighteen Dothideomycetes fungi. *PLoS Pathogens,* 8, e1003037.
OJEDA, D. I., DHILLON, B., TSUI, C. K. & HAMELIN, R. C. 2014. Single-nucleotide polymorphism discovery in *Leptographium longiclavatum*, a mountain pine beetle-associated symbiotic fungus, using whole-genome resequencing. *Molecular Ecology Resources,* 14, 401-10.
ÖKMEN, B. & DE WIT, P. J. G. M. 2012. *Cladosporium fulvum*–Tomato Pathosystem: Fungal Infection Strategy and Plant Responses. *Molecular Plant Immunity.* Wiley-Blackwell.
OLYMPUS 2015. [Online]. Available: http://www.olympus-lifescience.com/en/. [Accessed 8th December 2015].
OSPINA-GIRALDO, M. D. & JONES, R. W. 2003. Characterization of the glucose-6-phosphate isomerase gene in *Phytophthora infestans* reveals the presence of multiple alleles. *Fungal Genetics and Biology,* 40, 197-206.

PARK, J., LEE, S., CHOI, J., AHN, K., PARK, B., PARK, J., KANG, S. & LEE, Y. H. 2008. Fungal cytochrome p450 database. *BMC Genomics,* 9.
PARLANGE, F., DAVERDIN, G., FUDAL, I., KUHN, M. L., BALESDENT, M. H., BLAISE, F., GREZES-BESSET, B. & ROUXEL, T. 2009. *Leptosphaeria maculans* avirulence gene *AvrLm4-7* confers a dual recognition specificity by the *Rlm4* and *Rlm7* resistance genes of oilseed rape, and circumvents *Rlm4*-mediated recognition through a single amino acid change. *Molecular Microbiology,* 71, 851-863.
PATEL, D. A., ZANDER, M., VAN DE WOUW, A. P., MASON, A. S., EDWARDS, D. & BATLEY, J. 2015. Population Diversity of *Leptosphaeria maculans* in Australia. *International Journal of Biology,* 7, 18-36.
PEDRAS, M. S. C. & TAYLOR, J. L. 1993. Metabolism of the phytoalexin brassinin by the blackleg fungus. *Journal of Natural Products,* 56, 731-738.
PETERSEN, T. N., BRUNAK, S., VON HEIJNE, G. & NIELSEN, H. 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature Methods,* 8, 785-786.
PLISSONNEAU, C., DAVERDIN, G., OLLIVIER, B., BLAISE, F., DEGRAVE, A., FUDAL, I., ROUXEL, T. & BALESDENT, M. H. 2015. A game of hide and seek between avirulence genes *AvrLm4-7* and *AvrLm3* in *Leptosphaeria maculans*. *New Phytologist*. 209, 1613-24.
POLK, D. B. & PEEK, R. M. 2010a. *Helicobacter Pylori*: gastric cancer and beyond. *Nature Reviews Cancer,* 10, 593-593.
PURWANTARA, A., BARRINS, J. M., COZIJNSEN, A. J., ADES, P. K. & HOWLE, B. J. 2000. Genetic diversity of isolates of the *Leptosphaeria maculans* species complex from Australia, Europe and North America using amplified fragment length polymorphism analysis. *Mycological Research,* 104, 772-781.
PURWANTARA, A., SALISBURY, P. A., BURTON, W. A. & HOWLETT, B. J. 1998. Reaction of *Brassica juncea* (Indian mustard) lines to Australian isolates of *Leptosphaeria maculans* under glasshouse and field conditions. *European Journal of Plant Pathology,* 104, 895-902.
QIAGEN 2015. DNeasy Plant Mini Kit.
R CORE TEAM 2015. R: A Language and Environment for Statistical Computing. 3.0.2 ed. Vienna, Austria: R Foundation for Statistical Computing.
RABE, F., AJAMI-RASHIDI, Z., DOEHLEMANN, G., KAHMANN, R. & DJAMEI, A. 2013. Degradation of the plant defence hormone salicylic acid by the biotrophic fungus *Ustilago maydis*. *Molecular Microbiology,* 89, 179-88.
RAFFAELE, S. & KAMOUN, S. 2012. Genome evolution in filamentous plant pathogens: why bigger can be better. *Nature Reviews Microbiology,* 10, 417-30.
RAFIQI, M., GAN, P. H. P., RAVENSDALE, M., LAWRENCE, G. J., ELLIS, J. G., JONES, D. A., HARDHAM, A. R. & DODDS, P. N. 2010. Internalization of flax rust avirulence proteins into flax and tobacco cells can occur in the absence of the pathogen. *Plant Cell,* 22, 2017-2032.
RAMAN, R., TAYLOR, B., MARCROFT, S., STILLER, J., ECKERMANN, P., COOMBES, N., REHMAN, A., LINDBECK, K., LUCKETT, D., WRATTEN, N., BATLEY, J., EDWARDS, D., WANG, X. W. & RAMAN, H. 2012. Molecular mapping of qualitative and quantitative loci for resistance to *Leptosphaeria maculans* causing blackleg disease in canola (*Brassica napus* L.). *Theoretical and Applied Genetics,* 125, 405-418.
REMY, E., MEYER, M., BLAISE, F., CHABIRAND, M., WOLFF, N., BALESDENT, M. H. & ROUXEL, T. 2008a. The Lmpma1 gene of *Leptosphaeria maculans* encodes a plasma membrane H+-ATPase isoform essential for pathogenicity towards oilseed rape. *Fungal Genetics and Biology,* 45, 1122-1134.
REMY, E., MEYER, M., BLAISE, F., SIMON, U. K., KUHN, D., BALESDENT, M. H. & ROUXEL, T. 2009. A key enzyme of the leloir pathway is involved in pathogenicity of *Leptosphaeria maculans* toward oilseed rape. *Molecular Plant-Microbe Interactions,* 22, 725-736.

REMY, E., MEYER, M., BLAISE, F., SIMON, U. K., KUHN, D., CHABIRAND, M., RIQUELME, M., BALESDENT, M. H. & ROUXEL, T. 2008b. The *Lmgpi15* gene, encoding a component of the glycosylphosphatidylinositol anchor biosynthesis pathway, is required for morphogenesis and pathogenicity in *Leptosphaeria maculans*. *New Phytologist,* 179, 1105-1120.

ROUXEL, T. & BALESDENT, M. H. 2005. The stem canker (blackleg) fungus, *Leptosphaeria maculans*, enters the genomic era. *Molecular Plant Pathology,* 6, 225-241.

ROUXEL, T., CHUPEAU, Y., FRITZ, R., KOLLMANN, A. & BOUSQUET, J. F. 1988. Biological effects of sirodesmin-pl, a phytotoxin produced by *Leptosphaeria maculans*. *Plant Science,* 57, 45-53.

ROUXEL, T., GRANDAUBERT, J., HANE, J. K., HOEDE, C., VAN DE WOUW, A. P., COULOUX, A., DOMINGUEZ, V., ANTHOUARD, V., BALLY, P., BOURRAS, S., COZIJNSEN, A. J., CIUFFETTI, L. M., DEGRAVE, A., DILMAGHANI, A., DURET, L., FUDAL, I., GOODWIN, S. B., GOUT, L., GLASER, N., LINGLIN, J., KEMA, G. H. J., LAPALU, N., LAWRENCE, C. B., MAY, K., MEYER, M., OLLIVIER, B., POULAIN, J., SCHOCH, C. L., SIMON, A., SPATAFORA, J. W., STACHOWIAK, A., TURGEON, B. G., TYLER, B. M., VINCENT, D., WEISSENBACH, J., AMSELEM, J., QUESNEVILLE, H., OLIVER, R. P., WINCKER, P., BALESDENT, M.-H. & HOWLETT, B. J. 2011. Effector diversification within compartments of the *Leptosphaeria maculans* genome affected by Repeat-Induced Point mutations. *Nature Communications,* 2.

SALISBURY P.A., BALLINGER D.J., WRATTEN N., PLUMMER K.M., HOWLETT B.J. 1995. Blackleg disease on oilseed *Brassica* in Australia – a review. *Australian Journal of Experimental Agriculture*, 35, 665–672.

SARMA, G. N., MANNING, V. A., CIUFFETTI, L. M. & KARPLUS, P. A. 2005. Structure of Ptr ToxA: an RGD-containing host-selective toxin from *Pyrenophora tritici-repentis*. *Plant Cell,* 17, 3190-202.

SAUNDERS, D. G. O., WIN, J., CANO, L. M., SZABO, L. J., KAMOUN, S. & RAFFAELE, S. 2012. Using Hierarchical Clustering of Secreted Protein Families to Classify and Rank Candidate Effectors of Rust Fungi. *PloS One,* 7, e29847.

SELKER, E. U. 1990. Premeiotic instability of repeated sequences in *Neurospora crassa*. *Annual Review of Genetics,* 24, 579-613.

SELKER, E. U., CAMBARERI, E. B., JENSEN, B. C. & HAACK, K. R. 1987. Rearrangement of Duplicated DNA in Specialized Cells of *Neurospora*. *Cell,* 51, 741-752.

SHEN, J., ARAKI, H., CHEN, L., CHEN, J. Q. & TIAN, D. 2006. Unique evolutionary mechanism in *R* genes under the presence/absence polymorphism in *Arabidopsis thaliana*. *Genetics,* 172, 1243-50.

SHIN, J. H., BLAY, S., MCNENEY, B. & GRAHAM, J. 2006. LDheatmap: An R function for graphical display of pairwise linkage disequilibria between Single Nucleotide Polymorphisms. *Journal of Statistical Software,* 16 c03.

SINGHANIA, R. R., PATEL, A. K., SUKUMARAN, R. K., LARROCHE, C. & PANDEY, A. 2013. Role and significance of beta-glucosidases in the hydrolysis of cellulose for bioethanol production. *Bioresource Technology,* 127, 500-507.

SKAMNIOTI, P., FURLONG, R. F. & GURR, S. J. 2008a. Evolutionary history of the ancient cutinase family in five filamentous Ascomycetes reveals differential gene duplications and losses and in *Magnaporthe grisea* shows evidence of sub- and neo-functionalization. *New Phytologist,* 180, 711-21.

SKAMNIOTI, P., FURLONG, R. F. & GURR, S. J. 2008b. The fate of gene duplicates in the genomes of fungal pathogens. *Communicative and Integrative Biology,* 1, 196-198.

SMIT, A. F. A., HUBLEY, R. & GREEN, P. 2013-2015. *RepearMasker Open-4.0* [Online]. Available: http://www.repeatmasker.org [Accessed 13th December 2015].

SOANES, D. M., CHAKRABARTI, A., PASZKIEWICZ, K. H., DAWE, A. L. & TALBOT, N. J. 2012. Genome-wide transcriptional profiling of appressorium development by the rice blast fungus *Magnaporthe oryzae*. *Plos Pathogens,* 8.

SPANU, P. D., ABBOTT, J. C., AMSELEM, J., BURGIS, T. A., SOANES, D. M., STÜBER, K., LOREN VAN THEMAAT, E. V., BROWN, J. K. M., BUTCHER, S. A., GURR, S. J., LEBRUN, M.-H., RIDOUT, C. J., SCHULZE-LEFERT, P., TALBOT, N. J., AHMADINEJAD, N., AMETZ, C., BARTON, G. R., BENJDIA, M., BIDZINSKI, P., BINDSCHEDLER, L. V., BOTH, M., BREWER, M. T., CADLE-DAVIDSON, L., CADLE-DAVIDSON, M. M., COLLEMARE, J., CRAMER, R., FRENKEL, O., GODFREY, D., HARRIMAN, J., HOEDE, C., KING, B. C., KLAGES, S., KLEEMANN, J., KNOLL, D., KOTI, P. S., KREPLAK, J., LÓPEZ-RUIZ, F. J., LU, X., MAEKAWA, T., MAHANIL, S., MICALI, C., MILGROOM, M. G., MONTANA, G., NOIR, S., O'CONNELL, R. J., OBERHAENSLI, S., PARLANGE, F., PEDERSEN, C., QUESNEVILLE, H., REINHARDT, R., ROTT, M., SACRISTÁN, S., SCHMIDT, S. M., SCHÖN, M., SKAMNIOTI, P., SOMMER, H., STEPHENS, A., TAKAHARA, H., THORDAL-CHRISTENSEN, H., VIGOUROUX, M., WEßLING, R., WICKER, T. & PANSTRUGA, R. 2010. Genome expansion and gene loss in powdery mildew fungi reveal tradeoffs in extreme parasitism. *Science,* 330, 1543-1546.

SPERSCHNEIDER, J., DODDS, P. N., GARDINER, D. M., MANNERS, J. M., SINGH, K. B. & TAYLOR, J. M. 2015. Advances and challenges in computational prediction of effectors from plant pathogenic fungi. *PLoS Pathogens,* 11, e1004806.

SPERSCHNEIDER, J., GARDINER, D. M., TAYLOR, J. M., HANE, J. K., SINGH, K. B. & MANNERS, J. M. 2013. A comparative hidden Markov model analysis pipeline identifies proteins characteristic of cereal-infecting fungi. *BMC Genomics,* 14, 807.

STERGIOPOULOS, I. & DE WIT, P. J. 2009. Fungal effector proteins. *Annual Review of Phytopathology,* 47, 233-63.

STOTZ, H. U., MITROUSIA, G. K., DE WIT, P. J. G. M. & FITT, B. D. L. 2014. Effector-triggered defence against apoplastic fungal pathogens. *Trends in Plant Science,* 19, 491-500.

STRAUSS, J. & KUBICEK, C. P. 1990. Beta-glucosidase and cellulase formation by a *Trichoderma reesei* mutant defective in constitutive beta-glucosidase formation. *Journal of General Microbiology,* 136, 1321-1326.

STUART, L. M., PAQUETTE, N. & BOYER, L. 2013. Effector-triggered versus pattern-triggered immunity: how animals sense pathogens. *Nature Reviews Immunology,* 13, 199-206.

SUZUKI, R. & SHIMODAIRA, H. 2006. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. 22, 1540-1542.

TAKARA CLONTECH. *EmeraldAmp Max PCR Master Mix* [Online]. Available: http://catalog.takara-bio.co.jp/product/basic_info.php?unitid=U100006935 [Accessed 11th December 2015].

THERMOFISHER SCIENTIFIC 2015a. DNA Gel Loading Dye (6X).

THERMOFISHER SCIENTIFIC 2015b. GeneRuler™ DNA ladder.

TINDALL, E. A., PETERSEN, D. C., NIKOLAYSEN, S., MILLER, W., SCHUSTER, S. C. & HAYES, V. M. 2010. Interpretation of custom designed Illumina genotype cluster plots for targeted association studies and next-generation sequence validation. *BMC Research Notes,* 3, 39.

TIWARI, P., MISRA, B. N. & SANGWAN, N. S. 2013. Beta-Glucosidases from the Fungus *Trichoderma*: an efficient cellulase machinery in biotechnological applications. *Biomed Research International*, Article ID 203735.

TOLLENAERE, C., SUSI, H., NOKSO-KOIVISTO, J., KOSKINEN, P., TACK, A., AUVINEN, P., PAULIN, L., FRILANDER, M. J., LEHTONEN, R. & LAINE, A. L. 2012. SNP design from 454 sequencing of *Podosphaera plantaginis* transcriptome reveals a genetically diverse

pathogen metapopulation with high levels of mixed-genotype infection. *PLoS One,* 7, e52492.
TRAVADON, R., SACHE, I., DUTECH, C., STACHOWIAK, A., MARQUER, B. & BOUSSET, L. 2011. Absence of isolation by distance patterns at the regional scale in the fungal plant pathogen *Leptosphaeria maculans*. *Fungal Biology,* 115, 649-659.
VAN DE WOUW, A. P., COZIJNSEN, A. J., HANE, J. K., BRUNNER, P. C., MCDONALD, B. A., OLIVER, R. P. & HOWLETT, B. J. 2010. Evolution of linked avirulence effectors in *Leptosphaeria maculans* is affected by genomic environment and exposure to resistance genes in host plants. *PloS Pathogens,* 6.
VAN DE WOUW, A. P. & HOWLETT, B. J. 2012. Estimating frequencies of virulent isolates in field populations of a plant pathogenic fungus, *Leptosphaeria maculans*, using high-throughput pyrosequencing. *Journal of Applied Microbiology,* 113, 1145-1153.
VAN DE WOUW, A. P., LOWE, R. G., ELLIOTT, C. E., DUBOIS, D. J. & HOWLETT, B. J. 2014. An avirulence gene, *AvrLmJ1*, from the blackleg fungus, *Leptosphaeria maculans*, confers avirulence to *Brassica juncea* cultivars. *Molecular Plant Pathology,* 15, 523-30.
VAN DE WOUW, A. P. V., PETTOLINO, F. A., HOWLETT, B. J. & ELLIOTT, C. E. 2009. Mutations to *LmIFRD* affect cell wall integrity, development and pathogenicity of the ascomycete *Leptosphaeria maculans*. *Fungal Genetics and Biology,* 46, 695-706.
VE, T., WILLIAMS, S. J., CATANZARITI, A. M., RAFIQI, M., RAHMAN, M., ELLIS, J. G., HARDHAM, A. R., JONES, D. A., ANDERSON, P. A., DODDS, P. N. & KOBE, B. 2013. Structures of the flax-rust effector *AvrM* reveal insights into the molecular basis of plant-cell entry and effector-triggered immunity. *Proceedings of the National Academy of Sciences of the United States of America,* 110, 17594-17599.
VON HEIJNE, G. 1990. The signal peptide. *The Journal of Membrane Biology,* 115, 195-201.
WANG, X., GRUS, W. E. & ZHANG, J. 2006. Gene losses during human origins. *PLoS Biology,* 4, e52.
WANG, Y., LU, J., CHEN, S., SHU, L., PALMER, R. G., XING, G., LI, Y., YANG, S., YU, D., ZHAO, T. & GAI, J. 2014. Exploration of presence/absence variation and corresponding polymorphic markers in soybean genome. *Journal of Integrative Plant Biology,* 56, 1009-19.
WAPINSKI, I., PFEFFER, A., FRIEDMAN, N. & REGEV, A. 2007. Natural history and evolutionary principles of gene duplication in fungi. *Nature,* 449, 54-U36.
WARNES, G., GORJANC, G., LEISCH, F. & MAN, M. 2012. genetics: Population genetics. R package version 1.3.7. ed.
WEST, J. S., KHARBANDA, P. D., BARBETTI, M. J. & FITT, B. D. L. 2001. Epidemiology and management of *Leptosphaeria maculans* (phoma stem canker) on oilseed rape in Australia, Canada and Europe. *Plant Pathology,* 50, 10-27.
WICKER, T., OBERHAENSLI, S., PARLANGE, F., BUCHMANN, J. P., SHATALINA, M., ROFFLER, S., BEN-DAVID, R., DOLEZEL, J., SIMKOVA, H., SCHULZE-LEFERT, P., SPANU, P. D., BRUGGMANN, R., AMSELEM, J., QUESNEVILLE, H., VAN THEMAAT, E. V. L., PAAPE, T., SHIMIZU, K. K. & KELLER, B. 2013. The wheat powdery mildew genome shows the unique evolution of an obligate biotroph. *Nature Genetics,* 45, 1092-1096.
WILLIAMS, C. 1992. *Natural Selection: Domains, Levels, and Challenges,* New York, Oxford University Press.
WIN, J., CHAPARRO-GARCIA, A., BELHAJ, K., SAUNDERS, D. G., YOSHIDA, K., DONG, S., SCHORNACK, S., ZIPFEL, C., ROBATZEK, S., HOGENHOUT, S. A. & KAMOUN, S. 2012. Effector biology of plant-associated organisms: concepts and perspectives. *Cold Spring Harbour Symposia on Quantitative Biology,* 77, 235-47.
XU, J. R. 2006. Fundamentals of fungal molecular population genetic analysis. *Current Issues in Molecular Biology,* 8, 75-89.

YU, J., CHANG, P. K., EHRLICH, K. C., CARY, J. W., BHATNAGAR, D., CLEVELAND, T. E., PAYNE, G. A., LINZ, J. E., WOLOSHUK, C. P. & BENNETT, J. W. 2004. Clustered pathway genes in aflatoxin biosynthesis. *Applied Environmental Microbiology,* 70, 1253-62.

ZANDER, M. 2015. *Understanding the structure and variation within the genome of the pathogenic ascomycete fungus Leptosphaeria maculans.*

ZANDER, M., ALAMERY, S., LORBERG, S. K., PATEL, D. A., TOLLENAERE, R., EDWARDS, D. & BATLEY J. 2015. Understanding effector-triggered immunity of plants to pathogens: a genomics approach. *Plant Biotechnology Journal*, [Submitted].

ZANDER, M., PATEL, D. A., VAN DE WOUW, A., LAI, K., LORENC, M. T., CAMPBELL, E., HAYWARD, A., EDWARDS, D., RAMAN, H. & BATLEY, J. 2013. Identifying genetic diversity of avirulence genes in *Leptosphaeria maculans* using whole genome sequencing. *Functional and Integrative genomics,* 13, 295-308.

ZHANG, L. M., LUO, H., LIU, Z. Q., ZHAO, Y., LUO, J. C., HAO, D. Y. & JING, H. C. 2014. Genome-wide patterns of large-size presence/absence variants in sorghum. *Journal of Integrative Plant Biology,* 56, 24-37.

ZHANG, Z. M., ZHANG, X., ZHOU, Z. R., HU, H. Y., LIU, M., ZHOU, B. & ZHOU, J. 2013. Solution structure of the Magnaporthe oryzae avirulence protein *AvrPiz-t*. *Journal of Biomolecular NMR,* 55, 219-23.

ZHAO, Q. X., YUAN, S., ZHANG, Y. L., ZHU, H., DAI, C. C., YANG, F. & HAN, F. M. 2007. Expression, purification and characterization of pectate lyase A from *Aspergillus nidulans* in *Escherichia coli*. *World Journal of Microbiology & Biotechnology,* 23, 1057-1064.

ZHOU, R. & LINZ, J. E. 1999. Enzymatic function of the nor-1 protein in aflatoxin biosynthesis in *Aspergillus parasiticus*. *Applied and Environmental Microbiology,* 65, 5639-5641.